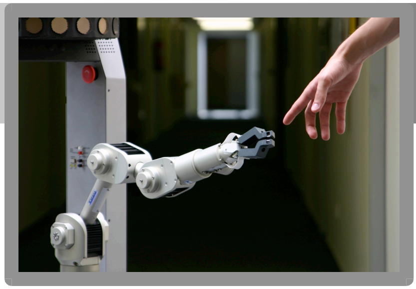
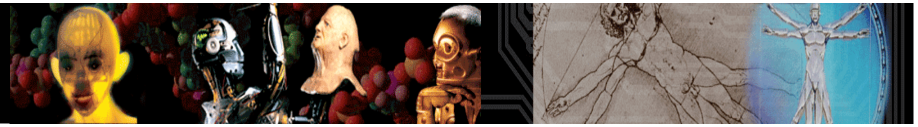




# Efficient CCG parsing of Spoken Dialogue For Human-Robot Interaction

Pierre Lison
Geert-Jan M. Kruijff

**Language Technology Lab**
**DFKI GmbH, Saarbrücken**
http://talkingrobots.dfki.de

Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

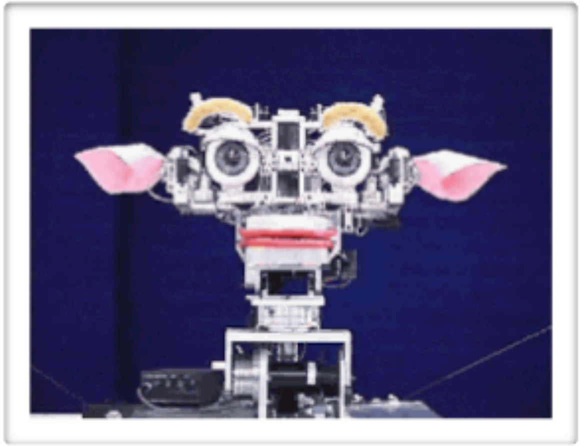
# Outline of the talk

- What is human-robot interaction?
- Overview of the approach
- Implementation
  - Robust CCG parsing of spoken dialogue
  - Incremental chart pruning
- Conclusion

# Outline of the talk

- **What is human-robot interaction?**
- Overview of the approach
- Implementation
  - Robust CCG parsing of spoken dialogue
  - Incremental chart pruning
- Conclusion

# What is human-robot interaction?

- Communication in all its aspects
  - Verbal- and non-verbal behaviours,
  - including gesture, posture, affective display, ...
  - at various interaction ranges (proximal, distant),
  - with reference to varying spatio-temporal contexts
- HRI in this talk
  - Focus on spoken dialogue, proximal interaction, varying spatial contexts

# Dialogue in HRI is (mostly) situated

Playing games on a table top...

Showing the robot around the “house“

Teaching the robot about new objects

Describing what kind of object it should be looking for, in some other location,

And trying to ask someone how to get to that location.

- **Situatedness** of spoken dialogue in HRI
  - Spoken dialogue in our case is often referential to aspects of the environment
  - “The environment” may refer to *small-scale space*, e.g. a table top, an area we are in,
  - But may also concern *large-scale space*, going beyond what is currently visible.
- Exploiting situatedness in **processing dialogue**
  - How to prime dialogue comprehension on the basis of situated context?

# Some problems to tackle

- The “usual” for spoken dialogue in HRI
  - Just like human spoken dialogue, dialogue in HRI is rife with incomplete or incorrect utterances, self-corrections, etc.
  - Pervasiveness of speech recognition errors
  - Ambiguities can arise at all processing levels
  - Extra-grammaticality (“out-of-coverage”) in relatively free dialogue

- Real-time dialogue system → strong **performance requirements**
  - The dialogue system must be capable of responding *quickly* to any utterance, even in the presence of noisy, ambiguous, or distorted input

    → Parsing must be **incremental**: (Partial) semantic interpretations should be constructed as soon as the first word is recognised, and be gradually extended as the utterance unfolds

    → Need to ensure the number of analyses remains **bounded** at each incremental processing step

# Disfluencies in spoken dialogue

- Extract from a corpus of task-oriented spoken dialogue : *The Apollo Lunar Surface Journal*.

***Parker*** : That’s all we need. Go ahead and park on your 045 **<okay>**. We’ll give you an update when you’re done.

***Cernan*** : Jack is **[it]** worth coming right there ?

***Schmitt*** : **err** looks like a pretty **go/** good location.

***Cernan*** : okay.

***Schmitt*** : We can sample the rim materials of this crater. **(Pause)** Bob, I’m at the **uh** south **uh** let’s say east-southeast rim of a, **oh**, 30-meter crater - **err** in the light mantle, of course - up on the **uh** Scarp and maybe 300...**(correcting himself) err** 200 meters from the **uh** rim of Lara in **(inaudible)** northeast direction.

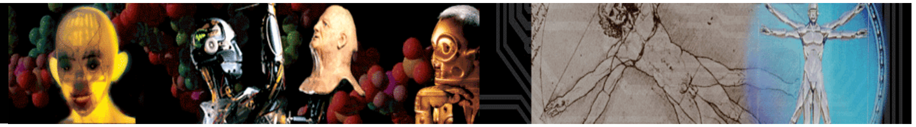
[ Play sound file ]

# Psycholinguistic motivation

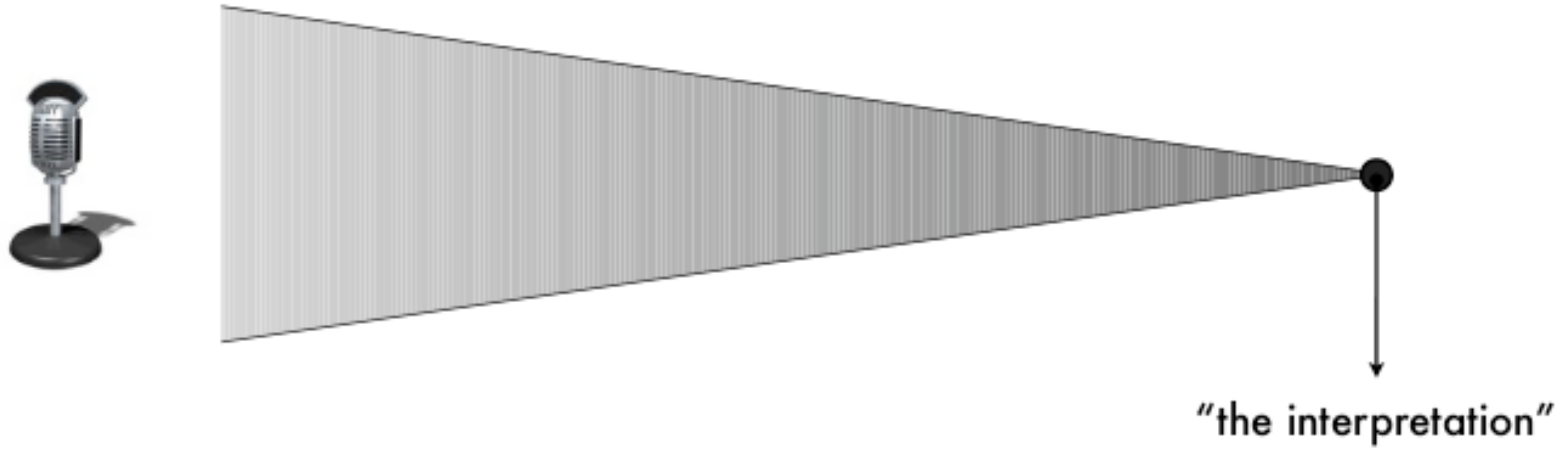
- How can we implement robust & efficient parsing of such noisy, ambiguous, distorted spoken inputs?
- Draw inspiration from how *humans* process dialogue
  - In visually situated dialogue, there is a close (bidirectional) *coupling* between how humans understand what they see, and what they hear
  - We know that this coupling is **closely time-locked,** as evidenced by
    - Empirical analyses of saccadic eye movements in visual scenes [Knoeferle & Crocker, 2006]
    - … and by neuroscience-based studies of event-related brain potentials (ERPs) [Van Berkum 2004]

→ At each processing step, **exploit the situated context** to predict, select, refine, extend, complement the interpretations, and increase parsing robustness

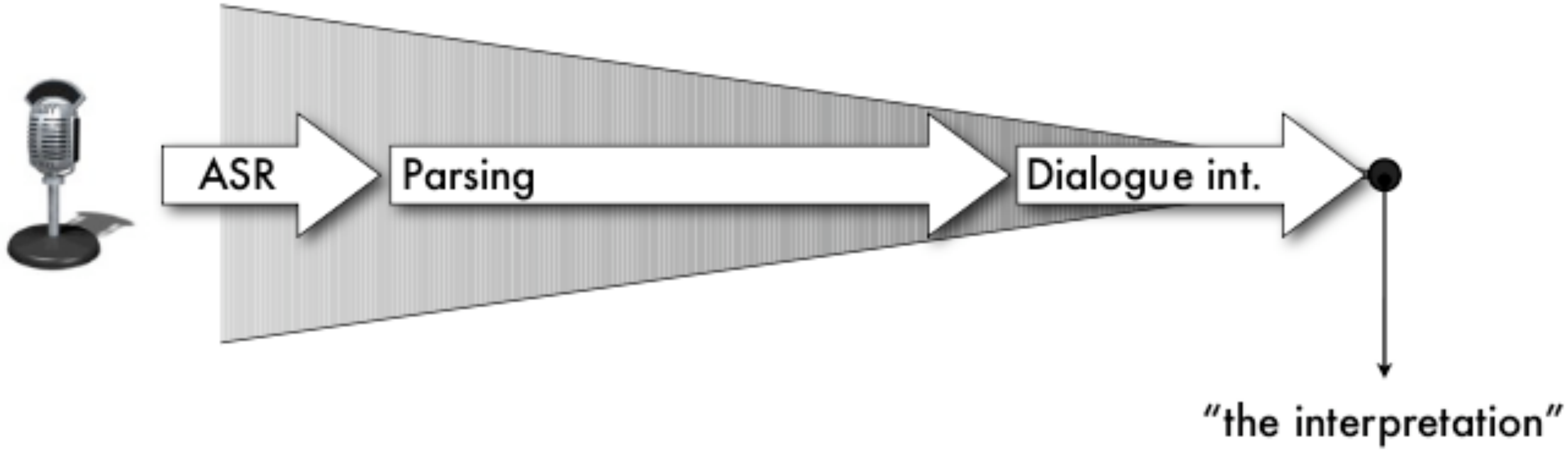
# Outline of the talk



- What is human-robot interaction?
- **Overview of the approach**
- Implementation
  - Robust CCG parsing of spoken dialogue
  - Improving parsing efficiency via incremental chart pruning
- Conclusion

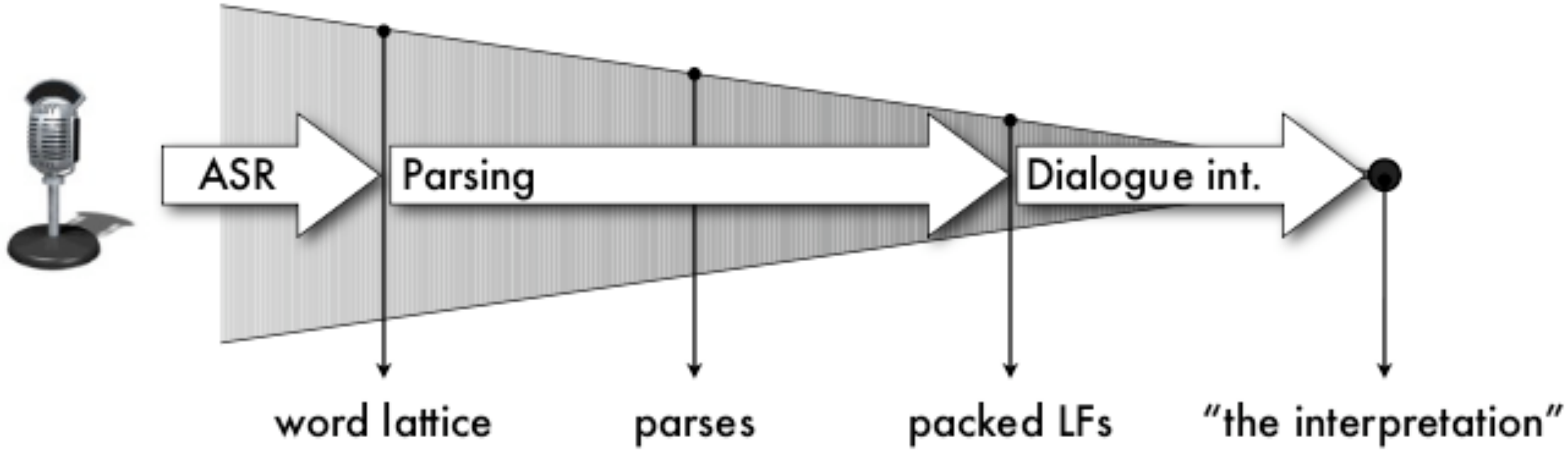

"the interpretation"

# Overview of the approach



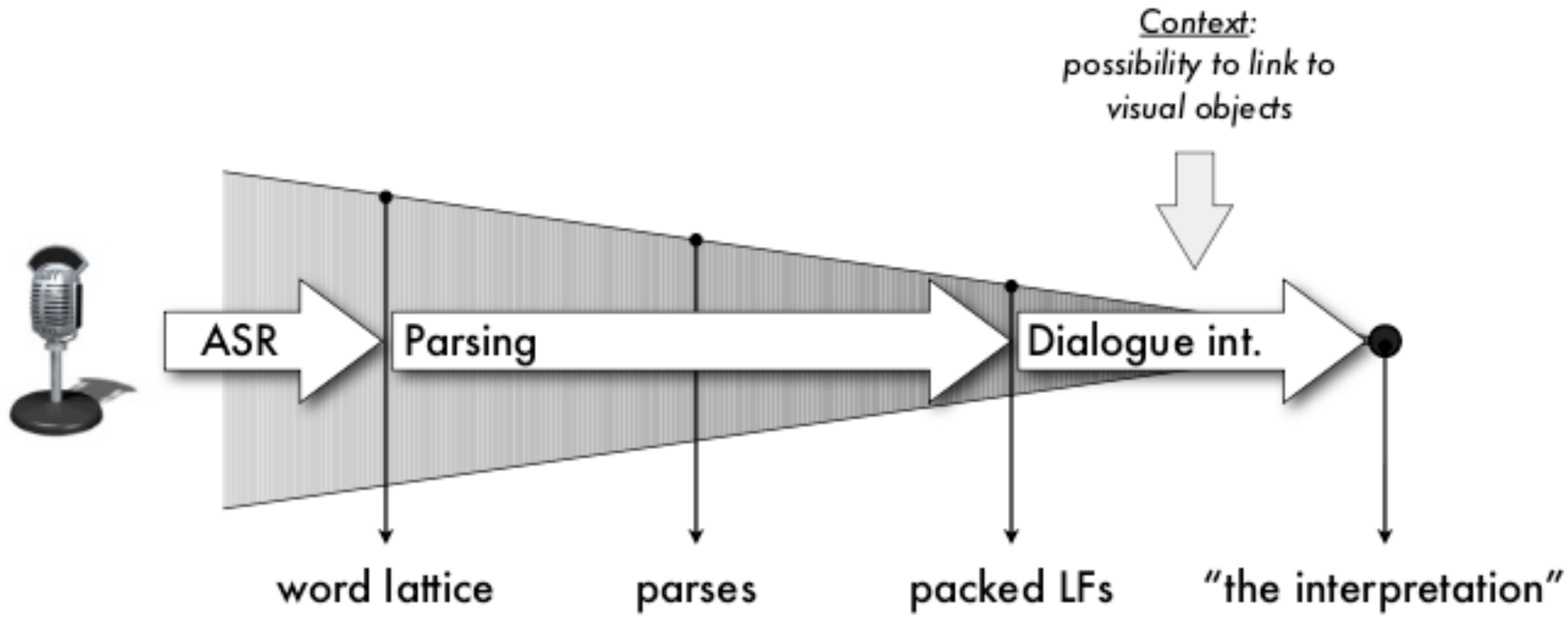

- Speech recognition with off-the-shelf ASR system
  - Language model is a class-based trigram statistical model
- Incremental parsing with Combinatory Categorial Grammar
- Dialogue interpretation tasks: reference resolution, dialogue move recognition, event structure interpretation

# Overview of the approach



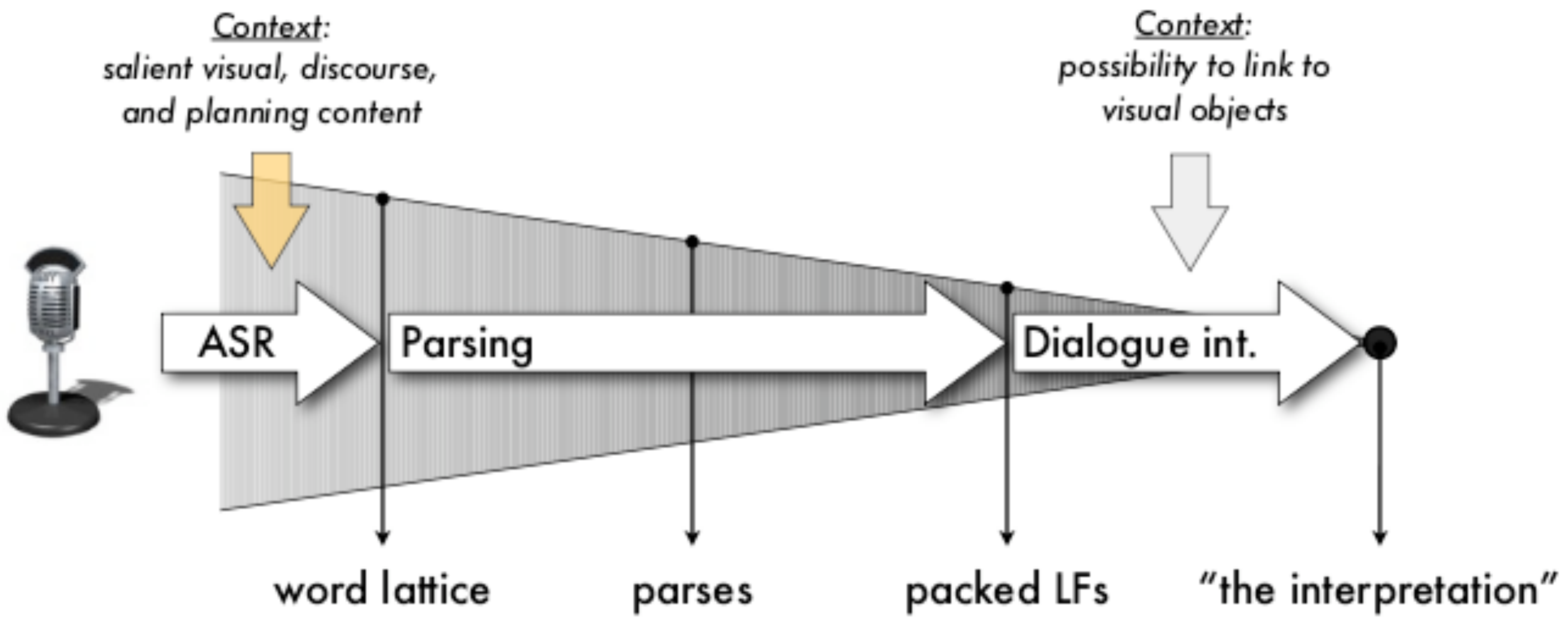

- Speech recognition outputs a *word lattice*
  - Word lattice = set of alternative recognition hypotheses compacted in a directed graph
- The CCG parser takes a word lattice as input and outputs packed logical forms, expressed in the HLDS formalism [Baldridge & Kruijff 2002]
  - Logical forms are ontologically rich, relational structures
- Dialogue interpretation based on a SDRT-like dialogue structure

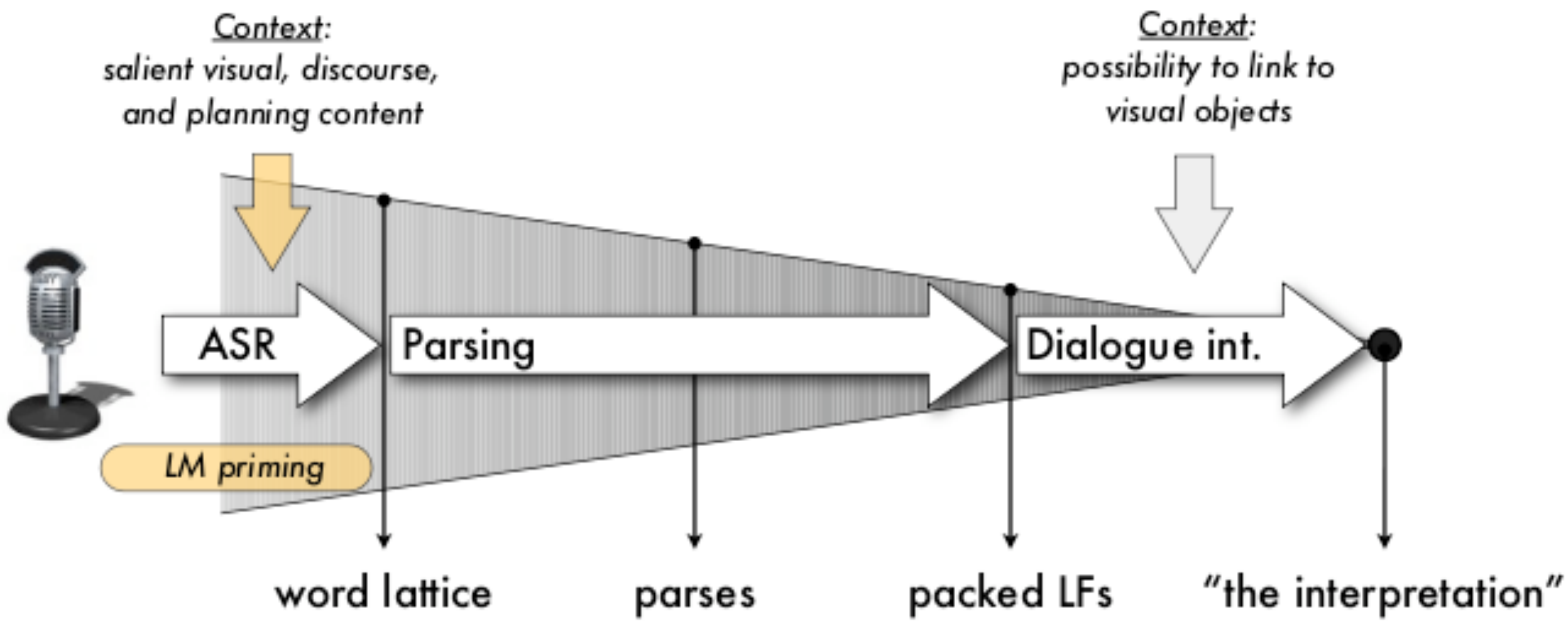
# Overview of the approach




- Linguistic interpretations must be associated with extra-linguistic knowledge about the environment
  - Dialogue needs to connect with other modalities like vision, spatial reasoning, navigation, manipulation, or planning.
- A specific module, called the "binder", is responsible for this cross-modal information binding (Ontology-based *mediation* across modalities)

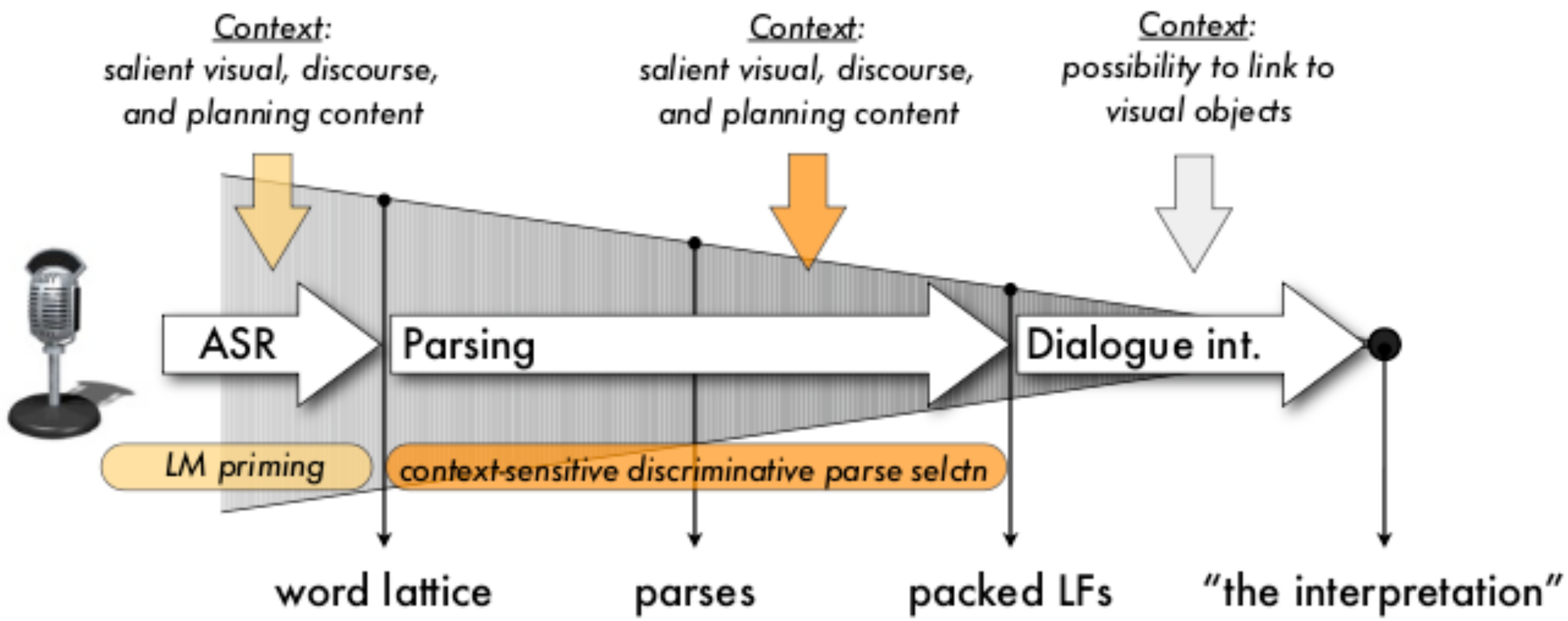
# Overview of the approach





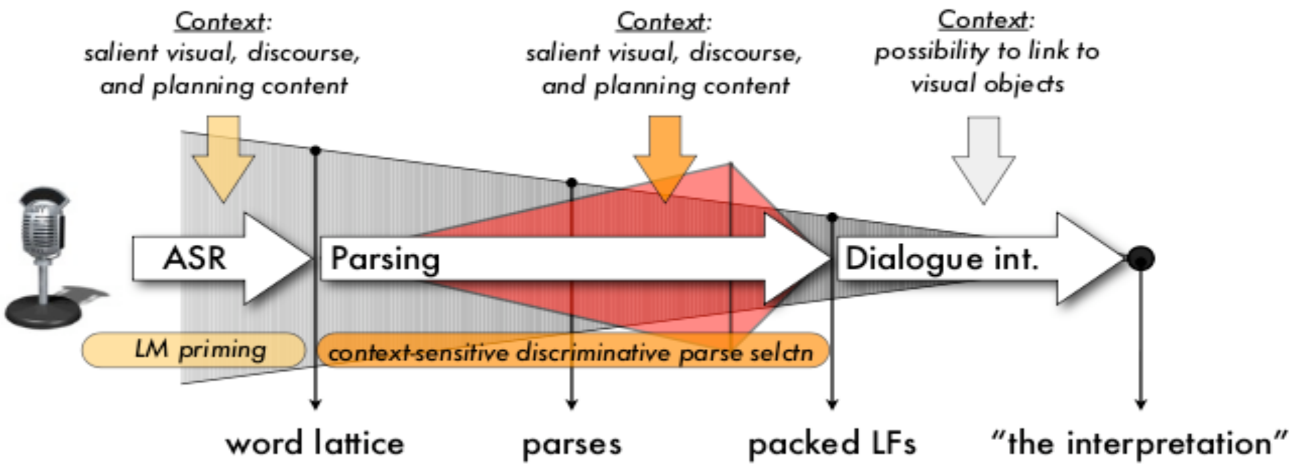

- Information about salient contextual entities are exploited to guide the speech recognition [Lison & Kruijff, 2008]
- *Objective*: establish expectations about uttered words which are most likely to be heard given the context

# Overview of the approach

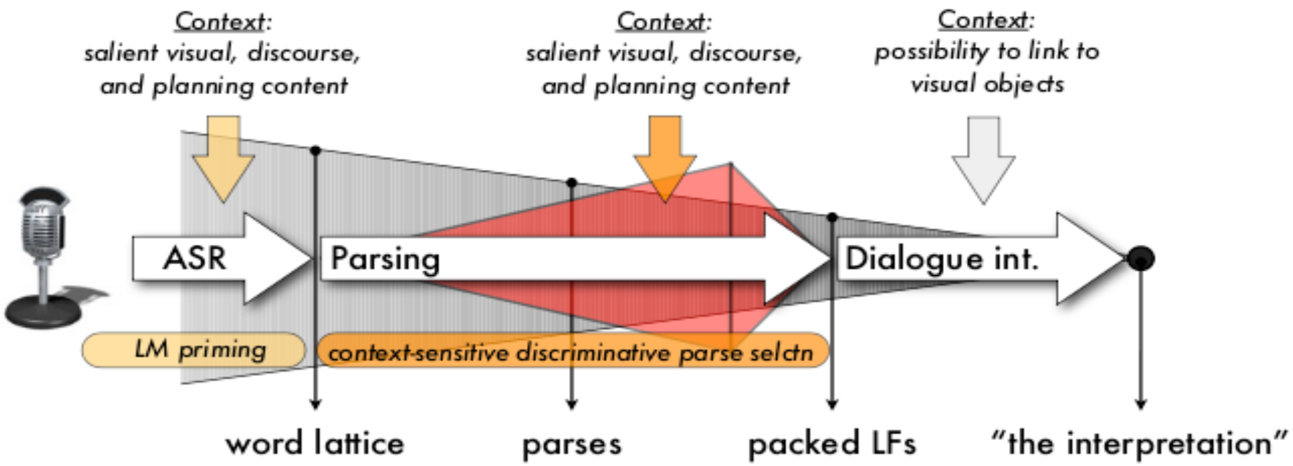



- Incremental parsing with Combinatory Categorial Grammar
  - Grammar able to handle ill-formed and misrecognised utterances by selectively relaxing and extending its set of grammatical rules.
- Use a (statistical) discriminative parse selection model to select the most likely parse(s) amongst the possible ones
  - Model includes various contextual features to guide the selection

# Overview of the approach




- Incremental integration of context, speech recognition and parsing
  - ASR processes the speech stream and outputs a word lattice
  - Incremental parsing processes the word lattice
  - The result is a packed representation of possible interpretations
  - … which is then incrementally filtered by the parse selection module to retain only the most likely partial interpretations




In three keywords:

- **Hybrid**: Combination of fined-grained linguistic resources with statistical models, able to deliver both *deep* and *robust* dialogue processing
- **Integrated**: goes all the way from the speech signal up to the semantic and pragmatic interpretation
- **Context-sensitive**: Context is used at every processing step to guide the comprehension, both an *anticipation* tool and a *discrimination* tool

# Outline of the talk

- What is human-robot interaction?
- Overview of the approach
- **Implementation**
  - Robust CCG parsing of spoken dialogue
  - Incremental chart pruning
- Conclusion

# Outline of the talk

- What is human-robot interaction?
- Overview of the approach
- Implementation
  - **Robust CCG parsing of spoken dialogue**
  - Incremental chart pruning
- Conclusion

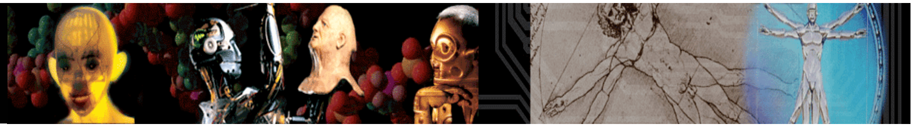
# Robust parsing of spoken dialogue

- Difficulty of parsing spoken input
  - Parsing needs to be robust to *ill-formed* and *misrecognized* input
  - Different approaches possible: Shallow parsing, statistical approaches, controlled relaxation of grammar rules
  - **Grammar relaxation** through non-standard CCG rules added in the grammar; inspired by [Zettlemoyer & Collins, 2007]

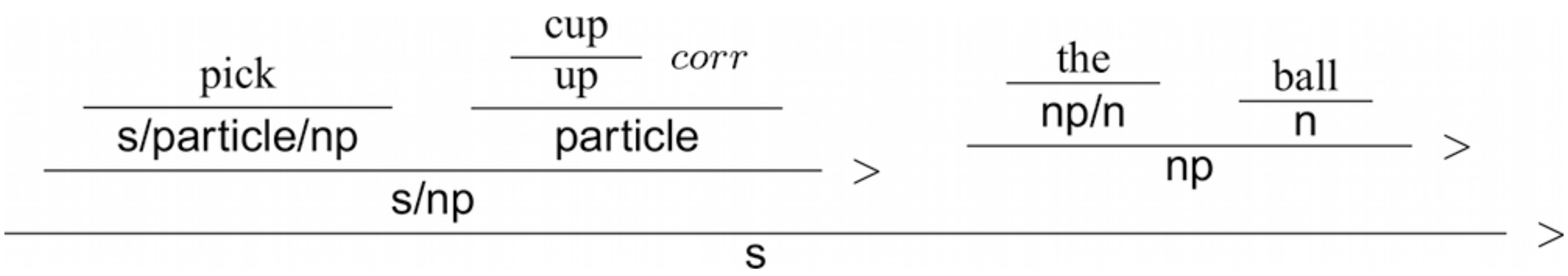
- Different types of rules:
  - *Type-shifting rules* to account for missing words
  - *"Paradigmatic heap"* rules for dealing with syntactic disfluencies
  - *Discourse-level composition rules* for combining discourse units
  - *ASR correction rules* for correcting misrecognized words

- Problem: better coverage and integration, but also more analyses

# Non-standard rules: example 1



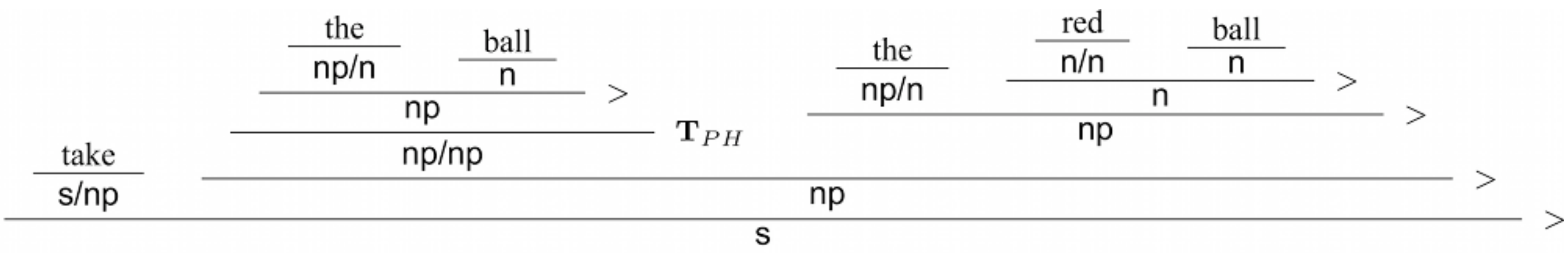
- Example of application of an ASR correction rule to accommodate a speech recognition error:




CCG derivation for "*Pick cup the ball*"

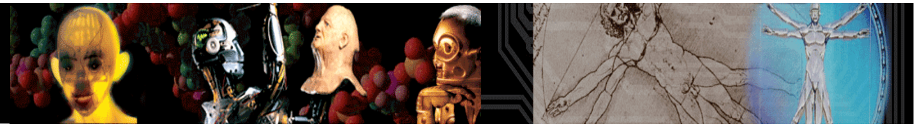
# Non-standard rules: example 2

- Example of application of a “paradigmatic heap” rule to accommodate a disfluency (in this case, a self-correction)

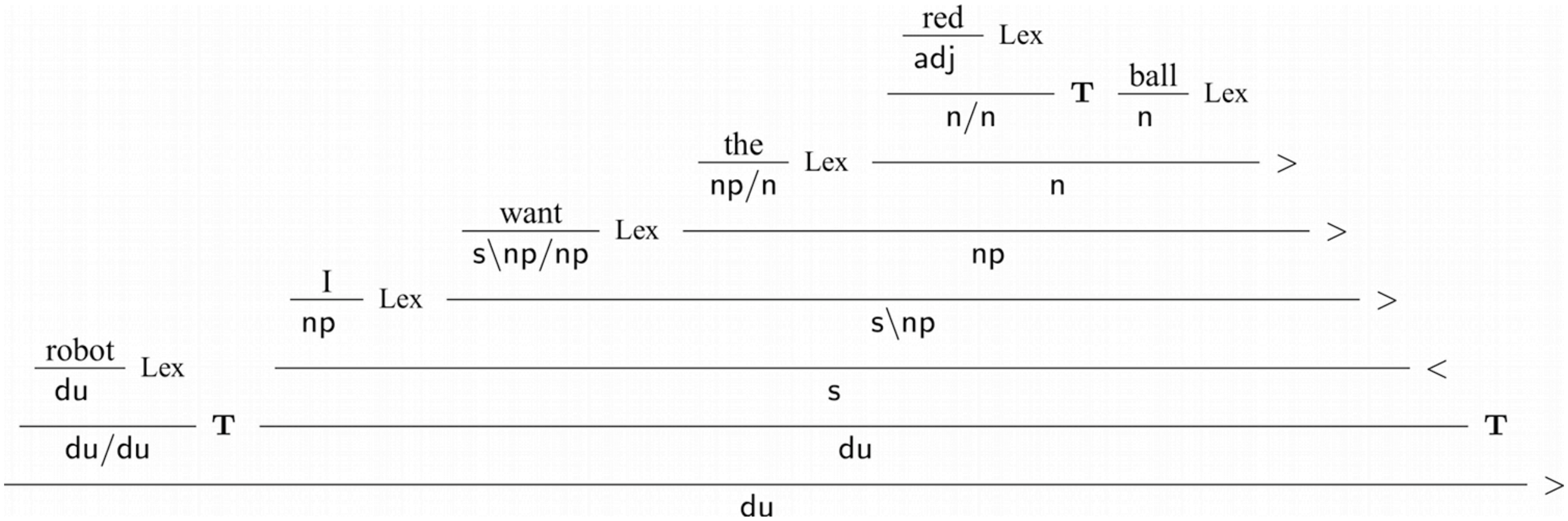



CCG derivation for “*Take the ball the red ball*”

# Non-standard rules: example 3



- Example of application of a discourse composition rule to combine discourse units




CCG derivation for “*Robot I want the red ball*”

# Discriminative parse selection

- Parsing produces a large number of analyses, arising from word lattice (multiple recognition hypotheses), controlled relaxation, and inherent ambiguity
- Need a mechanism to *filter/select* the resulting interpretations
- The task is represented as a function $F : X \rightarrow Y$ where the domain $X$ is the set of possible inputs (word lattices), and $Y$ the set of parses.
- The function $F$ , mapping a word lattice to its most likely parse, is then defined as :

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \ \mathbf{w}^T \cdot \mathbf{f}(x, y)$$

where $\mathbf{w}^T \cdot \mathbf{f}(x, y)$ is an inner product which can be seen as a measure of the "quality" of the parse.

# Parse selection features

- Given the parameters **w**, the optimal parse of a given word lattice *x* is determined by enumerating all parses generated by the grammar, extracting their features, computing the inner product $\mathbf{w}^T \cdot \mathbf{f}(x, y)$, and selecting the highest-scoring parse.

- Features include:
  - **acoustic features**: scores from speech recognition
  - **syntactic features**: derivational history of the parse
  - **semantic features**: substructures of the logical form
  - **contextual features**: situated and dialogue contexts

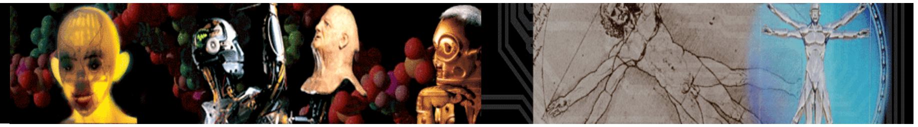
- The parameter vector **w** is learnt using a simple online perceptron

- Training on a corpus of automatically generated samples using a small domain-specific grammar

# Evaluation results

- Evaluation results show very significant improvements in accuracy and robustness over the baseline

  (see [Lison and Kruijff 2009] for details)



- What is human-robot interaction?
- Overview of the approach
- Implementation
  - Robust CCG parsing of spoken dialogue
  - **Improving parsing efficiency via incremental chart pruning**
- Conclusion

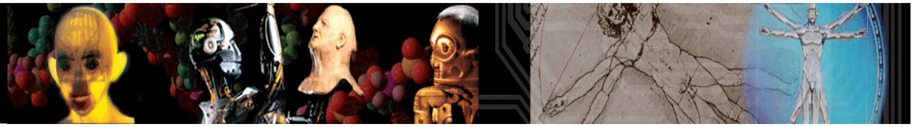
# Improving parsing efficiency

- Using parse selection *during* incremental parsing

- How does it work?
  - After each incremental parsing step, *rank* the partial analyses using the parse selection model (= score assignment)
  - Only keep a limited number of *high-scoring parses* in the parse chart
    - The exact number of parses to keep is determined by a beam width parameter ( optimal width ≈ 30 for our configuration)
  - All analyses outside the beam are *pruned* from the chart

- Effects of chart pruning
  - Chart pruning bounds parsing time and space complexity (which is crucial for real-time dialogue processing in HRI)

# Evaluation results

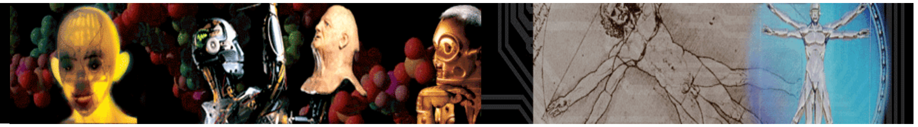


- We evaluated our incremental chart pruning mechanism on our Wizard-of-Oz corpus, with all grammar relaxation rules activated
    - Input: word lattices containing 10 recognition hypotheses
- Evaluations show statistically significant reductions in parsing time, with no large drop in accuracy (at least if the beam width parameter > 50 )

| | Beam width | Average parsing time per word lattice (in s.) | $F_1$-value for exact match | $F_1$-value for partial match |
|---|---|---|---|---|
| Baseline | (none) | **10.1** | 57.5 % | 89.8 % |
| | 120 | 5.78 | 57.5 % | 89.2 % |
| | 60 | **4.82** | 56.9 % | 87.4 % |
| | 40 | 4.66 | 54.9 % | 85.3 % |
| | 30 | 4.21 | 54.9 % | 84.2 % |

→ Empirical results on a WoZ corpus demonstrate a **53.8%** decrease in parsing time

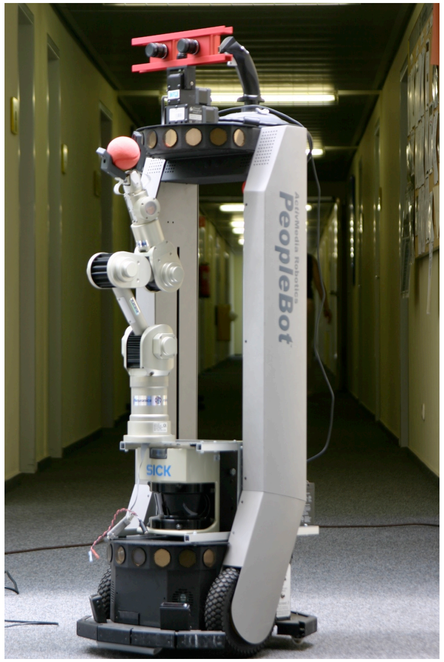
# Outline of the talk



- What is human-robot interaction?
- Overview of the approach
- Implementation
  - Robust CCG parsing of spoken dialogue
  - Improving parsing efficiency via incremental chart pruning
- **Conclusion**

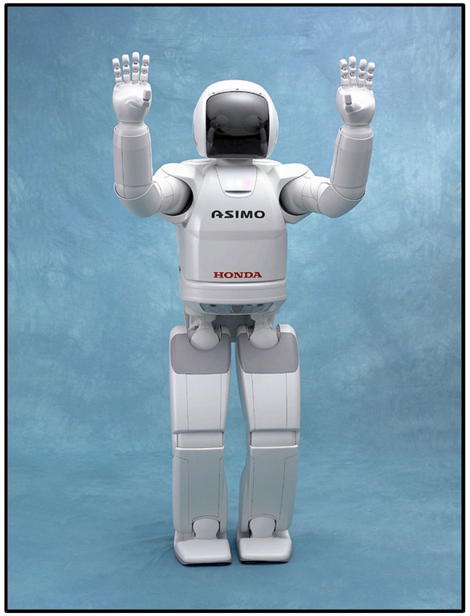
# Conclusions

- We presented an integrated, fully implemented approach to **situated spoken dialogue comprehension** for human-robot interaction
- Incremental parser based on *Combinatory Categorial Grammar*, taking word lattices as input, and outputting partial semantic interpretations
- Robust parsing of spoken inputs based on a *relaxed CCG grammar* coupled with a *discriminative model* exploring a wide range of linguistic and contextual features
- After each incremental parsing step, the partial semantic interpretations are filtered in order to retain only the most likely hypotheses in the chart
- **Forthcoming work:** use of more refined contextual features, extension of the grammar relaxation rules, experiments with more sophisticated machine learning algorithms, larger Wizard-of-Oz corpus




For more information, check our website:

*http://talkingrobots.dfki.de*




# Thanks for your attention!