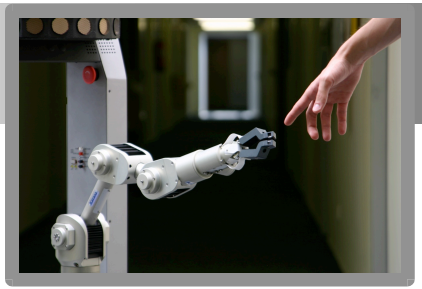
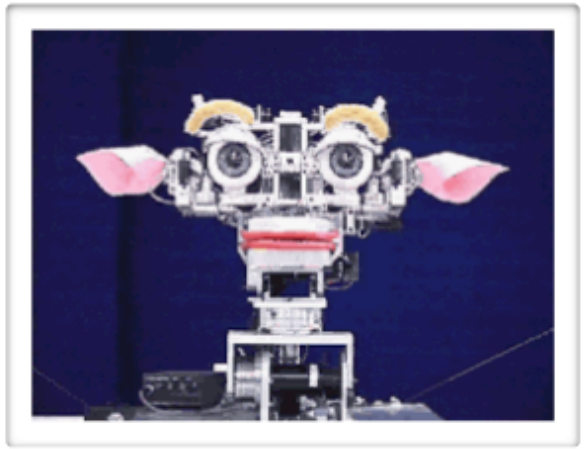


Robust Processing of Spoken Situated Dialogue

Pierre Lison
Geert-Jan M. Kruijff

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

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

Some problems to tackle

- The “usual” for spoken dialogue
  - Just like human spoken dialogue, dialogue in HRI is rife with partial, fragmentary, ungrammatical utterances, as well as many **disfluencies** (filled pauses, speech repairs, corrections, repetitions, etc.)
  - Pervasiveness of **speech recognition errors**
  - (+ ambiguity resolution, extra-grammaticality, etc.)

- Performance requirements for real-time dialogue
  - The system must be capable of responding *quickly* to any utterance, even in the presence of noisy, ambiguous, or distorted input

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*l*** 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.

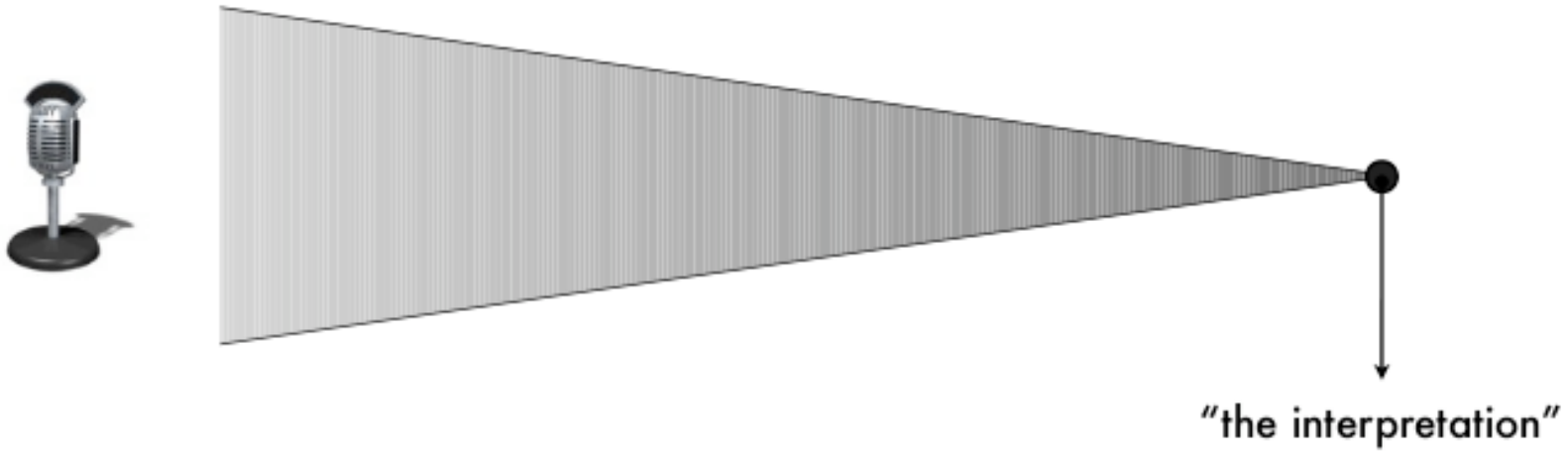
[Play sound file]

Psycholinguistic motivation

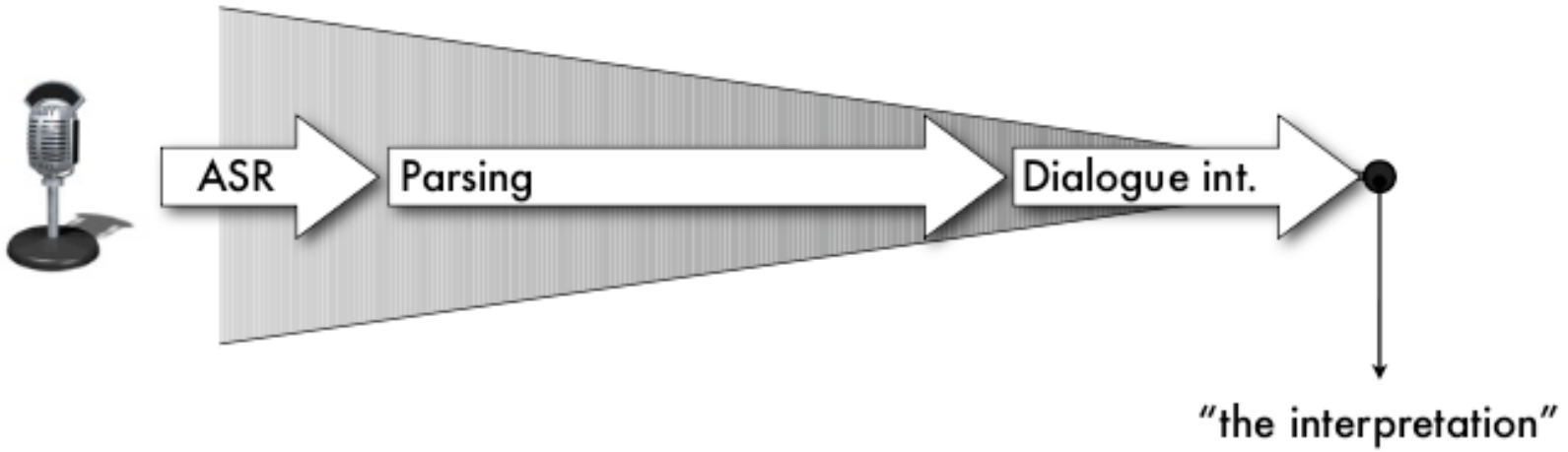
- How can we implement robust & efficient parsing of such noisy, ambiguous, distorted spoken inputs?
- Draw inspiration from how *humans* process dialogue
  - In visually situated dialogue, there is a close (bidirectional) *coupling* between how humans understand what they see, and what they hear
  - 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

Overview of the approach

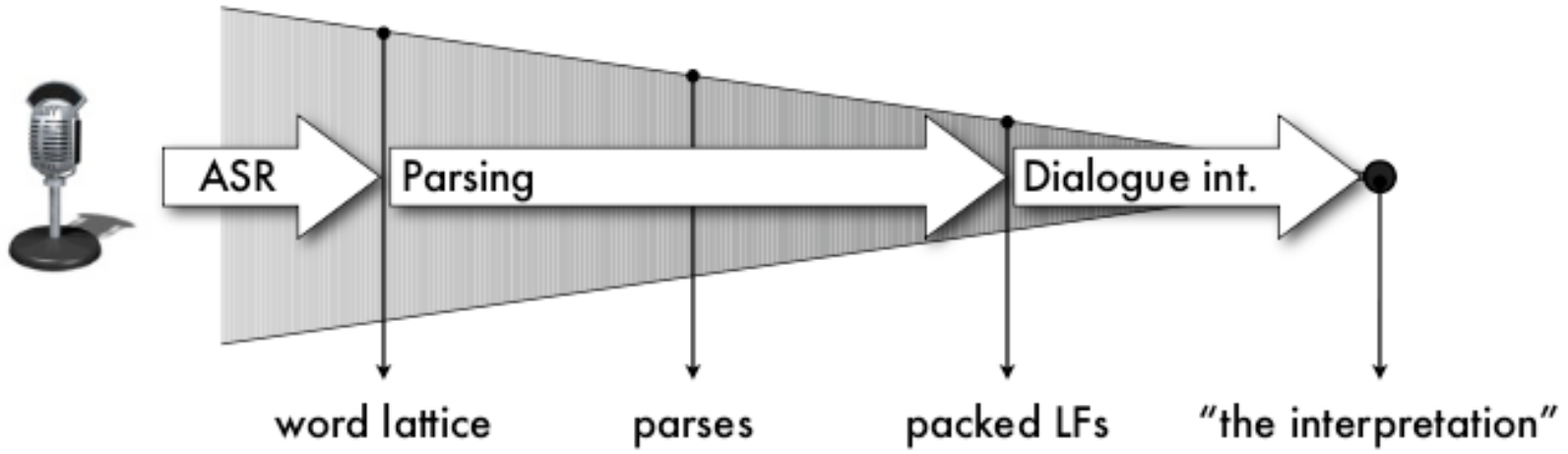


Overview of the approach



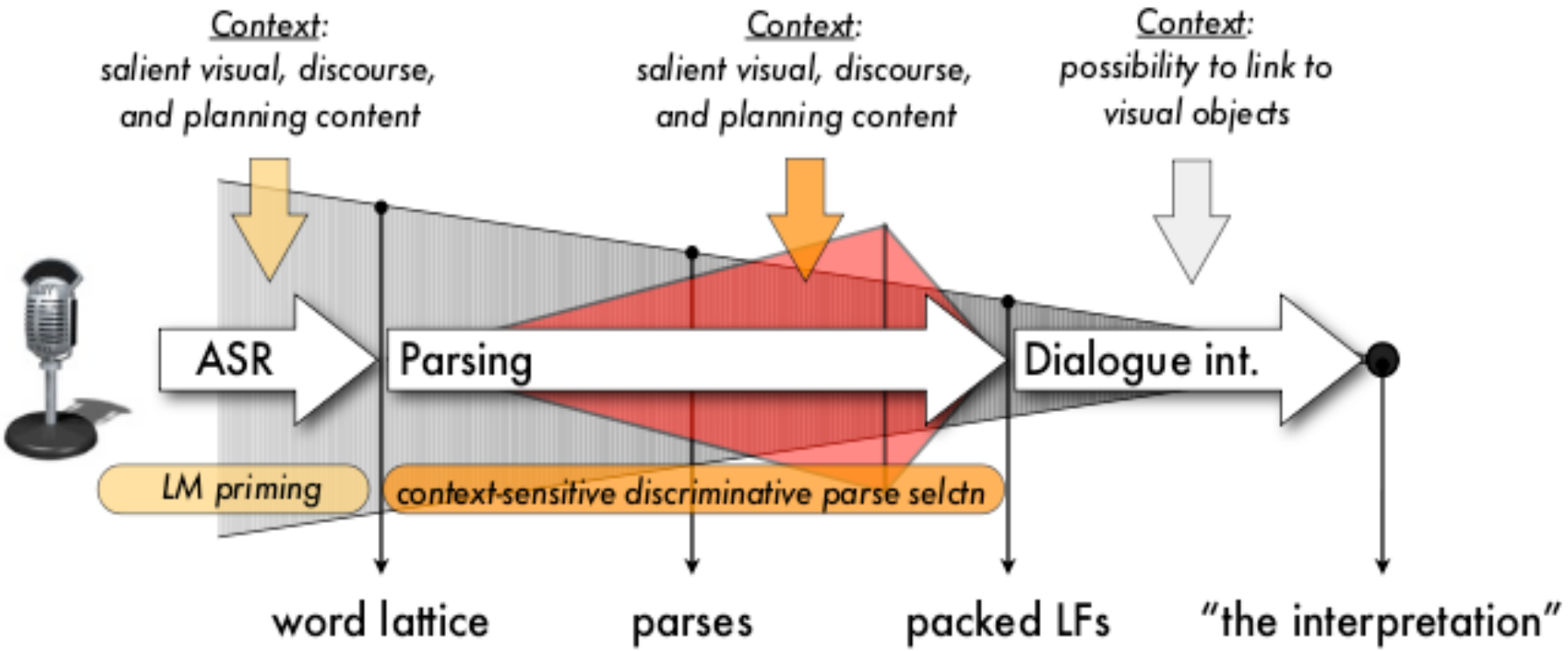
- Speech recognition with statistical models
- Incremental parsing with *Combinatory Categorial Grammar*
- Dialogue interpretation tasks: reference resolution, dialogue move recognition, etc.

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 semantic representations (logical forms)
  - Logical forms are ontologically rich, relational structures
- Dialogue interpretation based on dialogue structure

Overview of the approach



Robust parsing of spoken dialogue

- How to make parsing robust to both *ill-formed* and *misrecognised* spoken inputs ?

  → **Grammar relaxation** through the introduction of non-standard rules into the grammar

- Different types of rules to handle syntactic disfluencies (repetitions, corrections), hypothesise missing words, combine discourse units, and correct speech recognition errors

- Problem: better coverage, but also more analyses

  → Need a mechanism to *filter/select* the resulting interpretations

  → This is realised via a **parse selection** algorithm

Parse selection

- The parse selection is implemented via a discriminative statistical model including a broad range of linguistic and contextual features

- 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 discriminative model is trained using a simple online *perceptron*

Evaluation results

- We performed a quantitative evaluation of our approach, using its implementation on the fully integrated system
- *Testing data*: small Wizard-of-Oz corpus of human-robot interactions in a shared visual scene

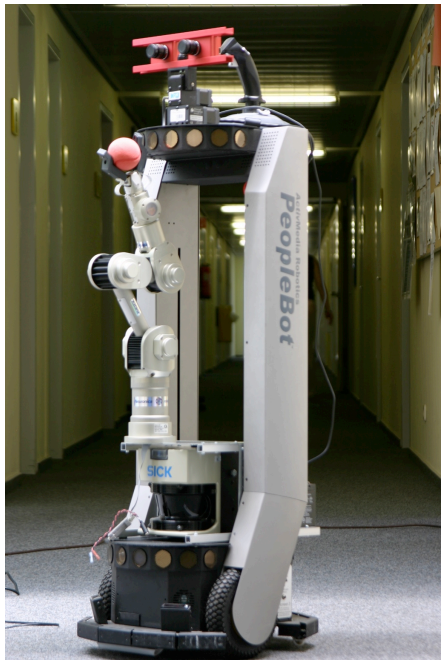
→ Evaluation results demonstrate significant improvements both in *accuracy* and *robustness* over the baseline:

- Relative increase of **55.6 %** for exact-match results (F_1 score)
- **27.6 %** for partial-match (also F_1 score)
- Decrease in Word Error Rate: from **20.5 %** to **15.7 %**

(see [Lison and Kruijff 2009] for details)

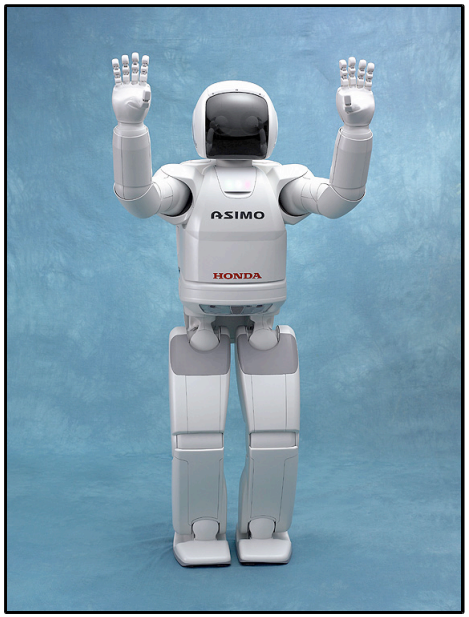
Conclusions

- We presented an integrated, fully implemented approach to **situated spoken dialogue comprehension** for human-robot interaction
- Incremental parser takes *word lattices* as input and outputs *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



For more information, check our website:

http://talkingrobots.dfki.de



Thanks for your attention!