A Salience-driven Approach to Speech Recognition for Human-Robot Interaction

Pierre Lison

Language Technology Lab,
German Research Centre for Artificial Intelligence (DFKI GmbH),
Saarbrücken Germany
plison@dfki.de

Abstract. We present an implemented model for speech recognition in natural environments which relies on contextual information about salient entities to prime utterance recognition. The hypothesis underlying our approach is that, in situated human-robot interaction, speech recognition performance can be significantly enhanced by exploiting knowledge about the immediate physical environment and the dialogue history. To this end, visual salience (objects perceived in the physical scene) and linguistic salience (previously referred-to objects within the current dialogue) are integrated into a single cross-modal salience model. The model is dynamically updated as the environment evolves, and is used to establish expectations about uttered words which are most likely to be heard given the context. The update is realised by continously adapting the word-class probabilities specified in the statistical language model. The present article discusses the motivations behind our approach, describes our implementation as part of a distributed, cognitive architecture for mobile robots, and reports the evaluation results on a test suite.

1 Introduction

Recent years have seen increasing interest in service robots endowed with communicative capabilities. In many cases, these robots must operate in open-ended environments and interact with humans using natural language to perform a variety of service-oriented tasks. Developing cognitive systems for such robots remains a formidable challenge. Software architectures for cognitive robots are typically composed of several cooperating subsystems, such as communication, computer vision, navigation and manipulation skills, and various deliberative processes such as symbolic planners [1].

These subsystems are highly interdependent. Incorporating basic functionalities for dialogue comprehension and production is not sufficient to make a robot interact naturally in situated dialogues. Crucially, dialogue managers for human-robot interaction also needs to relate language, action and situated reality in a unified framework, and enable the robot to use its perceptual experience to continuously learn and adapt itself to the environment.

The first step in comprehending spoken dialogue is *automatic speech recognition* [ASR]. For robots operating in real-world noisy environments, and dealing with utterances pertaining to complex, open-ended domains, this step is particularly difficult and error-prone. In spite of continuous technological advances, the performance of ASR remains for most tasks at least an order of magnitude worse than that of human listeners [2].

One strategy for addressing this issue is to use context information to guide the speech recognition by percolating contextual constraints to the statistical language model [3]. In this paper, we follow this approach by defining a *context-sensitive language model* which exploits information about salient objects in the visual scene and linguistic expressions in the dialogue history to prime recognition. To this end, a *salience model* integrating both visual and linguistic salience is used to dynamically compute lexical activations, which are incorporated into the language model at runtime.

Our approach departs from previous work on context-sensitive speech recognition by modeling salience as inherently cross-modal, instead of relying on just one particular modality such as gesture [4], eye gaze [5] or dialogue state [3]. The Fuse system described in [6] is a closely related approach, but limited to the processing of object descriptions, whereas our system was designed from the start to handle generic situated dialogues.

The structure of the paper is as follows: in the next section we briefly introduce the software architecture in which our system has been developed. We then describe in Section 3 our approach, detailing the salience model, and explaining how it is exploited within the language model used for speech recognition. We finally present in Section 4 the empirical evaluation of our approach, followed in Section 5 by conclusions.

2 Background

The approach we present in this paper is fully implemented and integrated into a distributed cognitive architecture for autonomous robots (see [7]). The architecture is divided into a set of subsystems. Each subsystem consists of a number of processes, and a working memory. The processes can access sensors, effectors, and the working memory to share information within the subsystem.

The robot is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks. Fig. 1 illustrates the architecture for the communication subsystem.

Starting with speech recognition, we process the audio signal to establish a *word lattice* containing statistically ranked hypotheses about word sequences. Subsequently, parsing constructs grammatical analyses for the given word lattice. A grammatical analysis constructs both a syntactic analysis of the utterance, and a representation of its meaning. The analysis is based on an incremental chart parser[1] for Combinatory Categorial Grammar [8]. These meaning representations are ontologically richly sorted, relational structures, formulated in a

[1] Built using the OpenCCG API: http://openccg.sf.net

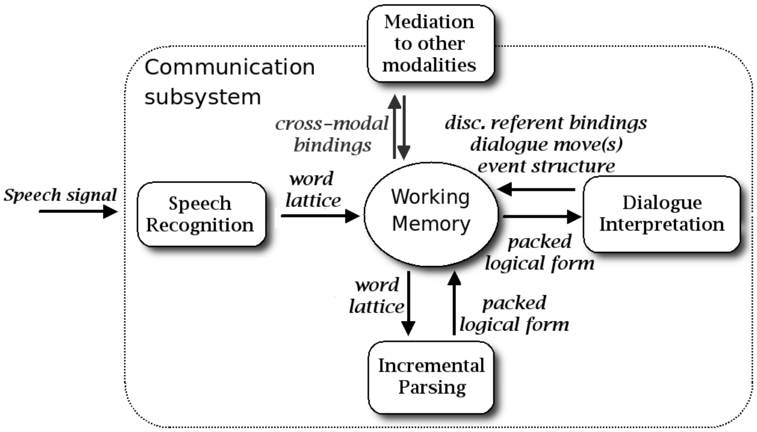

Fig. 1. Schema of the communication subsystem (limited to comprehension).

(propositional) description logic, more precisely in Hybrid Logic Dependency Semantics [9]. The parser then compacts all meaning representations into a single *packed logical form* [10, 11]. A packed logical form represents content similar across the different analyses as a single graph, using over- and underspecification of how different nodes can be connected to capture lexical and syntactic forms of ambiguity.

At the level of dialogue interpretation, the logical forms are resolved against a SDRS-like dialogue model [12], which is then exploited in various pragmatic interpretation tasks such as reference resolution or dialogue move recognition.

Linguistic interpretations must finally be associated with extra-linguistic knowledge about the environment – dialogue comprehension hence needs to connect with other subarchitectures like vision, spatial reasoning or planning. We realise this information binding between different modalities via a specific module, called the "binder", which is responsible for the ontology-based *mediation* accross modalities [13].

Interpretation *in context* indeed plays a crucial role in the comprehension of utterance as it unfolds. Human listeners continuously integrate linguistic information with scene understanding, (foregrounded entities and events) and word knowledge. This contextual knowledge serves the double purpose of interpreting what *has been* said, and predicting/anticipating what is *going to be* said. Their integration is also closely *time-locked*, as evidenced by analyses of saccadic eye movements in visual scenes [14] and by neuroscience-based studies of event-related brain potentials [15].

Several approaches in situated dialogue for human-robot interaction demonstrated that a robot's understanding can be substantially improved by relating utterances to the situated context [17, 18, 11]. Contextual knowledge can be fruitfully exploited to guide attention and help disambiguate and refine linguistic input by filtering out unlikely interpretations (see Fig. 2 for an illustration). Our

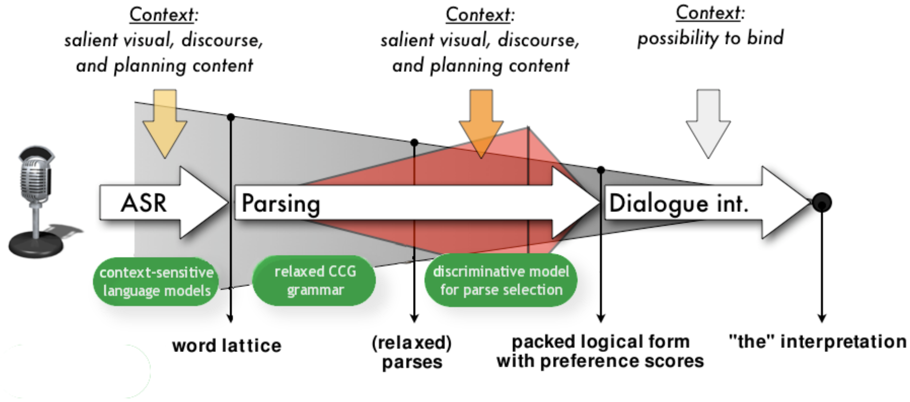
Context:
salient visual, discourse, and planning content
Context:
salient visual, discourse, and planning content
Context:
possibility to bind
ASR
Parsing
Dialogue int.
context-sensitive language models
relaxed CCG grammar
discriminative model for parse selection
word lattice
(relaxed) parses
packed logical form with preference scores
"the" interpretation

Fig. 2. Context-sensitivity in processing situated dialogue understanding (the use of contextual knowledge for discriminative parse selection is described in [16]).

approach is essentially an attempt to improve the speech recognition by drawing inspiration from the contextual priming effects evidenced in human cognition.

3 Approach

3.1 Salience modeling

In our implementation, we define *salience* using two main sources of information:

1. the salience of objects in the perceived visual scene;
2. the linguistic salience or "recency" of linguistic expressions in the dialogue history.

Other information sources could also be easily added in the model. Examples are the presence of gestures [4], eye gaze tracking [5], entities in large-scale space [19], or the integration of a task model – as salience generally depends on intentionality [20].

Visual salience Via the "binder", we can access the set of objects currently perceived in the visual scene. Each object is associated with a concept name (e.g. **printer**) and a number of features, for instance spatial coordinates or qualitative propreties like colour, shape or size.

Several features can be used to compute the salience of an object. The ones currently used in our implementation are (1) the object size and (2) its distance relative to the robot (i.e. spatial proximity). Other features could also prove to be helpful, like the reachability of the object or its distance from the point of visual focus – similarly to the spread of visual acuity across the human retina. To derive the visual salience value for each object, we assign a numeric value for the two variables, and then perform a weighted addition. The associated weights are determined via regression tests.

It is worth noting that the choice of a particular measure for the visual saliency is heavily dependent on the application domain and the properties of the visual scene (typical number of objects, relative distances, recognition capacities of the vision system, angle of view, etc.). For the application domain in which we performed our evaluation (cfr. section **??**), the experimental results turned out to be largely indifferent to the choice of a specific method of calculation for the visual saliency.

At the end of the processing, we end up with a set $\mathbf{E}_v$ of visual objects, each of which is associated with a numeric salience value $s(e_k)$, with $e_k \in \mathbf{E}_v$.

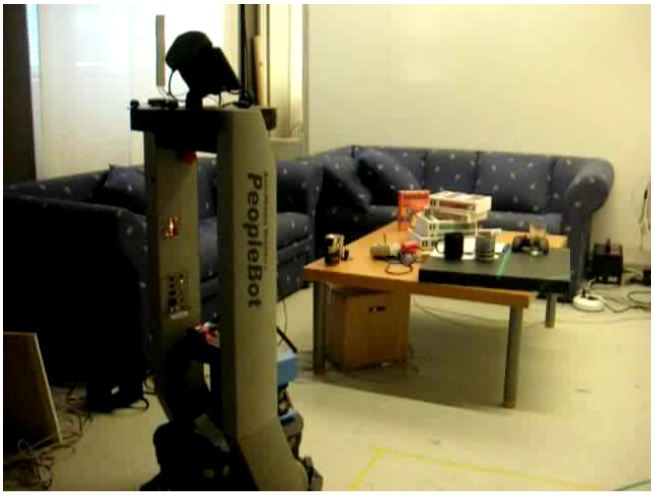

Fig. 3. Example of a visual scene

Linguistic salience There is a vast amount of literature on the topic of linguistic salience. Roughly speaking, linguistic salience can be characterised either in terms of *hierarchical recency*, according to a tree-like model of discourse structure (cfr. [21, 22, 12]), or in terms of *linear recency* of mention (see [23] for a discussion). Our implementation can theoretically handle both types of linguistic salience, but for all practical purposes, the system only takes linear recency into account, as it is easier to compute and usually more reliable than hierarchical recency (which crucially depends on having a well-formed discourse structure).

To compute the linguistic salience, we extract a set $\mathbf{E}_l$ of potential referents from the discourse structure, and for each referent e_k we assign a salience value

$s(e_k)$ equal to the distance (measured on a logarithmic scale) between its last mention and the current position in the discourse structure.

3.2 Cross-modal salience model

Once the visual and linguistic salience are computed, we can proceed to their integration into a *cross-modal statistical model*. We define the set $\mathbf{E}$ as the union of the visual and linguistic entities: $\mathbf{E} = \mathbf{E}_v \cup \mathbf{E}_l$, and devise a probability distribution $P(\mathbf{E})$ on this set:

$$P(e_k) = \frac{\delta_v \; I_{\mathbf{E}_v}(e_k) \; s_v(e_k) \quad + \quad \delta_l \; I_{\mathbf{E}_l}(e_k) \; s_l(e_k)}{|\mathbf{E}|} \tag{1}$$

where $I_A(x)$ is the indicator function of set A, and δ_v, δ_l are factors controlling the relative importance of each type of salience. They are determined empirically, subject to the following constraint to normalise the distribution :

$$\delta_v \sum_{e_k \in \mathbf{E}_v} s(e_k) + \delta_l \sum_{e_k \in \mathbf{E}_l} s(e_k) = |\mathbf{E}| \tag{2}$$

The statistical model $P(\mathbf{E})$ thus simply reflects the salience of each visual or linguistic entity: the more salient, the higher the probability.

3.3 Lexical activation

In order for the salience model to be of any use for speech recognition, a connection between the salient entities and their associated words in the ASR vocabulary needs to be established. To this end, we define a *lexical activation network*, which lists, for each possible salient entity, the set of words activated by it. The network specifies the words which are likely to be heard when the given entity is present in the environment or in the dialogue history. It can therefore include words related to the object denomination, subparts, common properties or affordances. The salient entity **laptop** will activate words like 'laptop', 'notebook', 'screen', 'opened', 'ibm', 'switch on/off', 'close', etc. The list is structured according to word classes, and a weight can be set on each word to modulate the lexical activation: supposing a **laptop** is present, the word 'laptop' should receive a higher activation than, say, the word 'close', which is less situation specific.

The use of lexical activation networks is a key difference between our model and [6], which relies on a measure of "descriptive fitness" to modify the word probabilities. One advantage of our approach is the possibility to go beyond object descriptions and activate word types denoting subparts, properties or affordances of objects. In the context of a **laptop** object, words such as 'screen', 'ibm', 'closed' or 'switch on/off' would for instance be activated.

If the probability of specific words is increased, we need to re-normalise the probability distribution. One solution would be to decrease the probability of all non-activated words accordingly. This solution, however, suffers from a significant drawback: our vocabulary contains many context-independent words like prepositions, determiners or general words like 'thing' or 'place', whose probability should remain constant. To address this issue, we mark an explicit distinction in our vocabulary between *context-dependent* and *context-independent* words. Only the context-dependent words can be activated or deactivated by the context. The context-independent words maintain a constant probability. Fig. 4 illustrates these distinctions.

In the current implementation, the lexical activation network is constructed semi-manually, using a simple lexicon extraction algorithm. We start with the list of possible salient entities, which is given by:

1. the set of physical objects the vision system can recognise ;
2. the set of nouns specified in the CCG lexicon with 'object' as ontological type.

For each entity, we then extract its associated lexicon by matching domain-specific syntactic patterns against a corpus of dialogue transcripts.

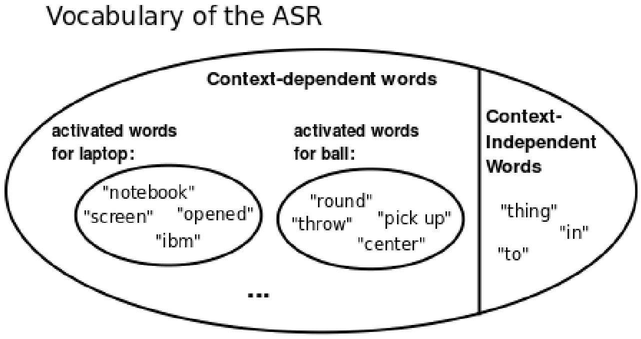

Fig. 4. Graphical illustration of the word activation network

3.4 Language modeling

We now detail the language model used for the speech recognition – a class-based trigram model enriched with contextual information provided by the cross-modal salience model.

3.5 Corpus generation

We need a corpus to train any statistical language model. Unfortunately, no corpus of situated dialogue adapted to our task domain is available to this day. Collecting in-domain data via Wizard of Oz experiments is a very costly and time-consuming process, so we decided to follow the approach advocated in [24] instead and generate a class-based corpus from a task grammar.

Practically, we first collected a small set of WoZ experiments, totalling about 800 utterances. This set is of course too small to be directly used as a corpus for language model training, but sufficient to get an intuitive idea of the utterances which are representative of our discourse domain. Based on it, we then designed a domain-specific context-free grammar able to cover most of the utterances. Weights were automatically assigned to each grammar rule by parsing our initial corpus, hence leading to a small *stochastic context-free grammar*.

As a last step, this grammar is randomly traversed a large number of times, which yields the final corpus.

3.6 Salience-driven, class-based language models

The objective of the speech recognizer is to find the word sequence $\mathbf{W}^*$ which has the highest probability given the observed speech signal $\mathbf{O}$ and a set $\mathbf{E}$ of salient objects:

$$\begin{aligned} \mathbf{W}^* &= \arg\max_{\mathbf{W}} \quad P(\mathbf{W}|\mathbf{O};\mathbf{E}) & (3) \\ &= \arg\max_{\mathbf{W}} \quad \underbrace{P(\mathbf{O}|\mathbf{W})}_{\text{acoustic model}} \times \underbrace{P(\mathbf{W}|\mathbf{E})}_{\text{salience-driven language model}} & (4) \end{aligned}$$

For a trigram language model, the probability of the word sequence $P(w_1^n|\mathbf{E})$ is:

$$P(w_1^n|\mathbf{E}) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2};\mathbf{E}) \tag{5}$$

Our language model is class-based, so it can be further decomposed into word-class and class transitions probabilities. The class transition probabilities reflect the language syntax; we assume they are independent of salient objects. The word-class probabilities, however, do depend on context: for a given class – e.g. *noun* -, the probability of hearing the word 'laptop' will be higher if a laptop is present in the environment. Hence:

$$P(w_i|w_{i-1}w_{i-2};\mathbf{E}) = \underbrace{P(w_i|c_i;\mathbf{E})}_{\text{word-class probability}} \times \underbrace{P(c_i|c_{i-1},c_{i-2})}_{\text{class transition probability}} \tag{6}$$

We now define the word-class probabilities $P(w_i|c_i;\mathbf{E})$:

$$P(w_i|c_i;\mathbf{E}) = \sum_{e_k \in \mathbf{E}} P(w_i|c_i,e_k) \times P(e_k) \tag{7}$$

To compute $P(w_i|c_i,e_k)$, we use the lexical activation network specified for e_k:

$$P(w_i|c_i, e_k) = \begin{cases} P(w_i|c_i) + \alpha_1 & \text{if} \quad w_i \in \text{activatedWords}(e_k) \\ P(w_i|c_i) - \alpha_2 & \text{if} \quad w_i \notin \text{activatedWords}(e_k) \wedge \\ & \qquad w_i \in \text{contextDependentWords} \\ P(w_i|c_i) & \text{else} \end{cases} \tag{8}$$

The optimum value of α_1 is determined using regression tests. α_2 is computed relative to α_1 in order to keep the sum of all probabilities equal to 1:

$$\alpha_2 = \frac{|\text{activatedWords}|}{|\text{contextDependentWords}| - |\text{activatedWords}|} \times \alpha_1$$

These word-class probabilities are dynamically updated as the environment and the dialogue evolves and incorporated into the language model at runtime.

4 Evaluation

4.1 Evaluation procedure

We evaluated our approach using a test suite of 250 spoken utterances recorded during Wizard-of-Oz experiments (a representative subset of the 800 utterances initially collected). The participants were asked to interact with the robot while looking at a specific visual scene. We designed 10 different visual scenes by systematic variation of the nature, number and spatial configuration of the objects presented. Fig. 5 gives an example of visual scene.

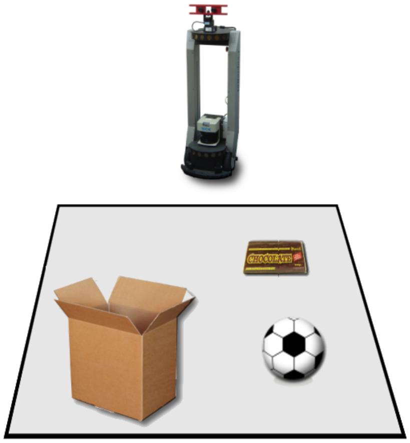

Fig. 5. Sample visual scene including three objects: a box, a ball, and a chocolate bar.

The interactions could include descriptions, questions and commands. No particular tasks were assigned to the participants. The only constraint we imposed was that all interactions with the robot had to be related to the shared visual scene.

After being recorded, all spoken utterances have been manually segmented one-by-one, and transcribed (without markers or punctuation).

4.2 Results

Table 1 summarises our experimental results. We focus our analysis on the WER of our model compared to the baseline – that is, compared to a class-based trigram model not based on salience.

Word Error Rate [WER]:	**Classical LM**	**Salience-driven LM**
vocabulary size *≃ 200 words*	25.04 % (NBest 3: 20.72 %)	24.22 % (NBest 3: 19.97 %)
vocabulary size *≃ 400 words*	26.68 % (NBest 3: 21.98 %)	**23.85** % (NBest 3: 19.97 %)
vocabulary size *≃ 600 words*	28.61 % (NBest 3: 24.59 %)	23.99 % (NBest 3: 20.27 %)

Table 1. Comparative results of recognition performance

The table details the WER results obtained by comparing the first recognition hypothesis to the gold standard transcription. Below these results, we also indicate the results obtained with NBest 3 – that is, the results obtained by considering the first three recognition hypotheses (instead of the first one). The word error rate is then computed as the *minimum* value of the word error rates yielded by the three hypotheses[2].

4.3 Analysis

As the results show, the use of a salience model can enhance the recognition performance in situated interactions: with a vocabulary of about 600 words,

[2] Or to put it slightly differently, the word error rate for NBest 3 is computed by assuming that, out of the three suggested recognition hypotheses, the one finally selected is always the one with the minimal error.

the WER is indeed reduced by $\frac{28.61-23.99}{28.61} \times 100 = \mathbf{16.1}$ % compared to the baseline. According to the *Sign* test, the differences for the last two tests (400 and 600 words) are statistically significant. As we could expect, the salience-driven approach is especially helpful when operating with a larger vocabulary, where the expectations provided by the salience model can really make a difference in the word recognition.

The word error rate remains nevertheless quite high. This is due to several reasons. The major issue is that the words causing most recognition problems are – at least in our test suite – function words like prepositions, discourse markers, connectives, auxiliaries, etc., and not content words. Unfortunately, the use of function words is usually not context-dependent, and hence not influenced by salience. By classifying the errors according to the part-of-speech of the misrecognised word, we estimated that 89 % of the recognition errors were due to function words. Moreover, our test suite is constituted of “free speech” interactions, which often include lexical items or grammatical constructs outside the range of our language model.

5 Conclusion

We have presented an implemented model for speech recognition based on the concept of *salience*. This salience is defined via *visual* and *linguistic* cues, and is used to compute degrees of *lexical activations*, which are in turn applied to dynamically adapt the ASR *language model* to the robot’s environment and dialogue state. The obtained experimental results demonstrate the effectiveness of our approach.

It is worth noting that the primary role of the context-sensitive ASR mechanism outlined in this paper is to establish *expectations* about uttered words which are most likely to be heard given the context – that is, to *anticipate* what will be uttered. In [16], we move a step further, and explain how we can also use the context as a *discrimination* tool to select the most relevant interpretations of a given utterance.

Acknowledgements

My thanks go to G.-J. Kruijff, H. Zender, M. Wilson and N. Yampolska for their insightful comments. The research reported in this article was supported by the EU FP6 IST Cognitive Systems Integrated project *Cognitive Systems for Cognitive Assistants* “CoSy” FP6-004250-IP. The first version of this paper appeared in [25].

References

1. Langley, P., Laird, J.E., Rogers, S.: Cognitive architectures: Research issues and challenges. Technical report, Institute for the Study of Learning and Expertise, Palo Alto, CA (2005)

2. Moore, R.K.: Spoken language processing: piecing together the puzzle. Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing **49** (2007) 418–435
3. Gruenstein, A., Wang, C., Seneff, S.: Context-sensitive statistical language modeling. In: Proceedings of INTERSPEECH 2005. (2005) 17–20
4. Chai, J.Y., Qu, S.: A salience driven approach to robust input interpretation in multimodal conversational systems. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing 2005, Vancouver, Canada, Association for Computational Linguistics (October 2005) 217–224
5. Qu, S., Chai, J.: An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In: Proceedings of the Conference of the North America Chapter of the Association of Computational Linguistics. (2007) 284–291
6. Roy, D., Mukherjee, N.: Towards situated speech understanding: visual context priming of language models. Computer Speech & Language **19**(2) (April 2005) 227–248
7. Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G.M., Brenner, M., Berginc, G., Skocaj, D.: Towards an integrated robot with multiple cognitive functions. In: AAAI, AAAI Press (2007) 1548–1553
8. Steedman, M., Baldridge, J.: Combinatory categorial grammar. In Borsley, R., Börjars, K., eds.: Nontransformational Syntax: A Guide to Current Models. Blackwell, Oxford (2009)
9. Baldridge, J., Kruijff, G.J.M.: Coupling CCG and hybrid logic dependency semantics. In: ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, Association for Computational Linguistics (2002) 319–326
10. Carroll, J., Oepen, S.: High efficiency realization for a wide-coverage unification grammar. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'05). (2005) 165–176
11. Kruijff, G., Lison, P., Benjamin, T., Jacobsson, H., Hawes, N.: Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In: Language and Robots: Proceedings from the Symposium (LangRo'2007), Aveiro, Portugal (December 2007) 55–64
12. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press (2003)
13. Jacobsson, H., Hawes, N., Kruijff, G.J., Wyatt, J.: Crossmodal content binding in information-processing architectures. In: Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Amsterdam, The Netherlands (March 12–15 2008)
14. Knoeferle, P., Crocker, M.: The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. Cognitive Science (2006)
15. Van Berkum, J.: Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things? In Carreiras, M., Jr., C.C., eds.: The on-line study of sentence comprehension: Eyetracking, ERPs and beyond. Psychology Press, New York NY (2004) 229–270
16. Lison, P.: Robust processing of situated spoken dialogue. In Chiarcos, C., de Castilho, R.E., Stede, M., eds.: Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Narr Verlag (2009) Proceedings of the Biennial GSCL Conference 2009 , Potsdam, Germany.

17. Roy, D.: Semiotic schemas: A framework for grounding language in action and perception. Artificial Intelligence **167**(1-2) (2005) 170–205
18. Brick, T., Scheutz, M.: Incremental natural language processing for HRI. In: Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07). (2007) 263 – 270
19. Zender, H., Kruijff, G.J.M.: Towards generating referring expressions in a mobile robot scenario. In: Language and Robots: Proceedings of the Symposium, Aveiro, Portugal (December 2007) 101–106
20. Landragin, F.: Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems. Signal Processing **86**(12) (2006) 3578–3595
21. Grosz, B.J., Sidner, C.L.: Attention, intentions, and the structure of discourse. Computational Linguistics **12**(3) (1986) 175–204
22. Grosz, B.J., Weinstein, S., Joshi, A.K.: Centering: a framework for modeling the local coherence of discourse. Computational Linguistics **21**(2) (1995) 203–225
23. Kelleher, J.: Integrating visual and linguistic salience for reference resolution. In Creaney, N., ed.: Proceedings of the 16th Irish conference on Artificial Intelligence and Cognitive Science (AICS-05), Portstewart, Northern Ireland (2005)
24. Weilhammer, K., Stuttle, M.N., Young, S.: Bootstrapping language models for dialogue systems. In: Proceedings of INTERSPEECH 2006, Pittsburgh, PA (2006)
25. Lison, P.: A salience-driven approach to speech recognition for human-robot interaction. In: Proceedings of the 13th ESSLLI student session, Hamburg (Germany) (2008)