Salience-driven Contextual Priming of Speech Recognition for Human-Robot Interaction

Pierre Lison and **Geert-Jan Kruijff** [1]

Abstract. The paper presents an implemented model for priming speech recognition, using contextual information about salient entities. The underlying hypothesis is that, in human-robot interaction, speech recognition performance can be improved by exploiting knowledge about the immediate physical situation and the dialogue history. To this end, visual salience (objects perceived in the physical scene) and linguistic salience (objects, events already mentioned in the dialogue) are integrated into a single cross-modal salience model. The model is dynamically updated as the environment changes. It is used to establish expectations about which words are most likely to be heard in the given context. The update is realised by continuously adapting the word-class probabilities specified in a statistical language model. The paper discusses the motivations behind the approach, and presents the implementation as part of a cognitive architecture for mobile robots. Evaluation results on a test suite show a statistically significant improvement of salience-driven priming speech recognition (WER) over a commercial baseline system.

1 Introduction

Service robots are becoming more and more sophisticated. In many cases, these robots must operate in open-ended environments and interact with humans using spoken natural language to perform a variety of service-oriented tasks. This has led to an increasing interest in developing dialogue systems for robots [28, 15, 23]. A fundamental challenge here is, how the robot can *situate* the dialogue: The robot should be able to understand what is being said, *and* how that relates to the physical situation [20, 25, 26, 11].

The relation between language and experience is often characterized as being *bi-directional* (cf. [14]). That is, language influences how to perceive the environment – and vice versa, the physical situation provides a context against which to interpret language. In this paper, we focus on how information from the dialogue- and situated context can help guiding, and improving, automatic speech recognition (ASR) in human-robot interaction (HRI). Spoken dialogue is one of the most natural means of communication for humans. Despite significant technological advances, however, ASR remains for most tasks at least an order of magnitude worse than that of human listeners [17]. This particularly holds for using ASR in HRI systems which typically have to operate in real-world noisy environments, dealing with utterances pertaining to complex, open-ended domains.

In this paper we present an approach to using context in priming ASR. By priming we mean, focusing the domain of words / word sequences ASR can expect next, so as to improve recognition. This approach has been implemented, and integrated into a cognitive architecture for a mobile robot [10, 14]. Evaluation results on a test suite with recordings of ”free speech” in the application domain show a statistically significant decrease in word-error rate (WER) of the implemented system, over a commercial baseline system.

We follow [9] and use context information (in the form of contextual constraints) to update the statistical language model used in ASR. We define 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. A *salience model* integrating both visual and linguistic salience [12] is used to dynamically compute lexical activations, which are incorporated into the language model at runtime.

The structure of the paper is as follows. We first situate our approach against the background of situated dialogue and ASR, and introduce the software architecture in which our system has been integrated. We then describe the salience model, and explain how it is utilised within the language model used for ASR. We finally present the evaluation of our approach, followed by conclusions.

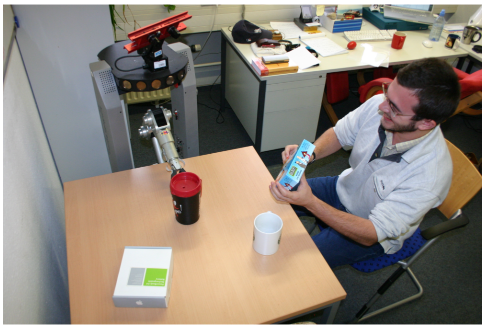

Figure 1. Example interaction

2 Background

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of utterance as it unfolds [13]. During utterance comprehension, humans combine linguistic information with scene understanding and ”world knowledge.”

Several approaches in processing situated dialogue for HRI have made similar observations [19, 20, 21, 4, 14]: A robot’s understanding can be improved by relating utterances to the situated context. This first of all presumes the robot *is able to* relate language and the world around it. [22] present a comprehensive overview of existing

[1] DFKI GmbH, Saarbrücken, Germany, email: {pierre.lison} {gj}@dfki.de

approaches. One of the earliest systems which connected utterances to a visual world was Winograd's SHRDLU [30]. Among more recent approaches, the most developed are those by Gorniak & Roy, and Steels *et al*. Gorniak & Roy [6, 7] present an approach in which utterance meaning is probabilistically mapped to visual and spatial aspects of objects in the current scene. Recently, they have extended their approach to include action-affordances [8]. Their focus has primarily been on relating language to the situation, and not on priming effects; same for SHRDLU. Steels *et al* [27, 25, 26] develop an approach where the connection between word meaning and percepts is modeled as a semiotic network, in which abstract categories mediate between language and the visual world.

Our approach on context-sensitive priming of speech recognition departs from previous work by modeling salience as inherently *cross-modal*, instead of relying on just one particular modality such as gesture [5], eye gaze [18] or dialogue state [9]. The FUSE system described in [21] is a closely related approach, but limited to the processing of object descriptions, whereas our system was designed from the start to handle generic dialogues. We can therefore prime not only words related to the object linguistic description, but also words denoting subparts, general properties, and affordances.

3 Architecture

Our approach has been implemented as part of a *distributed cognitive architecture* [10]. 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. Furthermore, across across subsystems can be interconnected (or "bound") [11]. Below, we first discuss ideas implemented in the comprehension side of the dialogue system, and then briefly point to several technical details. For more details, we refer to [14].

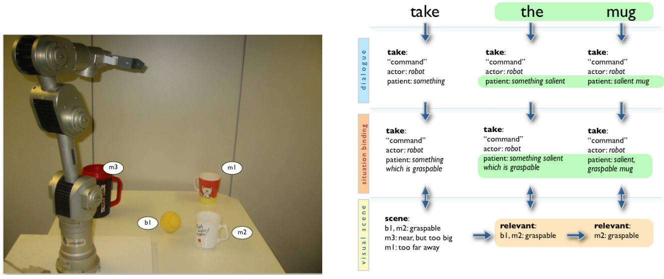

Figure 2. Dialogue comprehension: Visual scene (l.) and processing (r.)

Consider the visual scene in Fig. 2. There are three mugs, labelled $m1$ through $m3$, and a yellow ball $b1$. $b1$, $m2$ and $m3$ are in reach of the robot arm. The robot is only able to pick up $m2$ and $b1$ – $m3$ is too big, and $m1$ is too far away. The diagram in Fig. 2 illustrates (abstractly) what is happening in the dialogue system.

As the robot hears 'take', it starts parsing. All it can do at the moment is assuming it is dealing with a command (the utterance-initial verb indicates imperative mood), which it needs to execute. On linguistic grounds, it does not know what it is to manipulate. Grounding the command in the situation, the robot then applies categorical inferences, to interconnect linguistic content with content from other modalities. These inferences yield further information: a **take** action presupposes it applies to a *graspable* object. When the robot combines this information with what it knows about the visuo-spatial scene, it can already narrow down the set of *potential* objects to which the action may apply: Only $b1$ and $m2$ are graspable.

After processing 'the', the utterance meaning built up so far indicates the object should be salient. This constraint is met by $b1$ and $m2$. Finally, 'mug' completes the utterance. Using the narrowed down set of visual objects, the robot can determine $m2$ as the visual referent for the expression 'the mug' – even though, as a referring expression, 'the mug' is at best ambiguous between $m3$ and $m2$! Building up an interpretation of the utterance thus needs to be combined with categorical knowledge and information about the visuo-spatial scene. Only in this way the robot can arrive at an interpretation which "makes sense" in context.

Of course, the example in Fig. 2 is idealized. What if someone would have said 'it', instead of 'the mug'? To understand 'take it', the robot needs to be able to resolve the pronoun 'it' to a previously mentioned object. Furthermore, once it knows what object is meant, it should also be able to retrieve the corresponding visual referent. (In Fig. 2, the steps in green would have been affected.) Other, not unrealistic problems would have been if the vision system would not have been able to recognize any of the objects as mugs.This arises e.g. in scenarios for interactive visual learning. It can partly be resolved through categorical inference, which can establish that a 'mug' is a type of thing. Thus, even if the robot would only have recognized the ball, and discern between $m2$ and $m3$ in terms of graspability, it would have been able to resolve 'the mug' to $m2$. (In Fig. 2, the steps in orange would have been affected.) Finally, the utterance may have been ambiguous: 'Put the mug to the left of the ball to the right of the mug.' Which mug? And where is the robot supposed to move "the" mug? [4, 3, 14] discuss how such linguistic ambiguities can be resolved using the situated context.

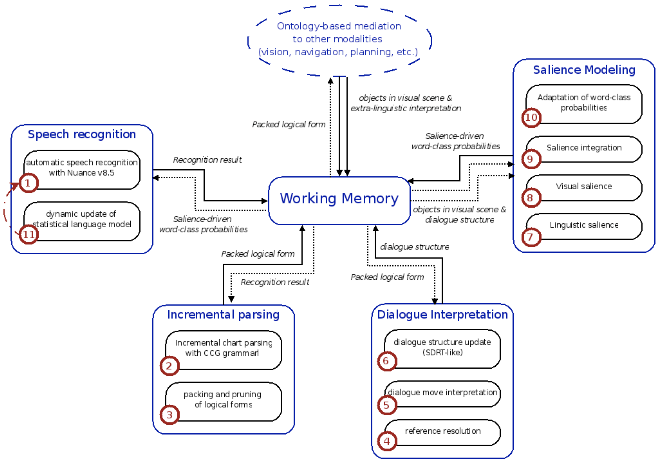

Figure 3. Schema of the system for spoken dialogue comprehension

Fig. 3 illustrates part of the implemented dialogue system. The first step in spoken dialogue comprehension is speech recognition. The ASR module is based on Nuance Recognizer v8.5 together with a statistical language model. For the online update of word class probabilities according to the salience model, we use the "just-in-time grammar" functionality provided by Nuance. The objective of the ASR module is to analyse the audio signal to form a *recognition result*, which is given as a word lattice with associated confidence levels. Once a (partial or complete) recognition result is available, it is added to the working memory.

The strings included in the word lattice are then parsed, to represent the syntactic and semantic structure of an utterance. Parsing only interprets an utterance at the grammatical level. Parsing is based on an incremental chart parser[2] for Combinatory Categorial Grammar [24, 1]. The parser yields a set of interpretations expressed as ontologically rich, relational structures (description logic-like; cf. [2]). These interpretations are *packed* into a single representation [14], a technique which enables us to efficiently handle ambiguities.

Dialogue interpretation tries to relate utterance meaning to the preceding context, by resolving contextual references, event structure, and dialogue moves (”speech acts”). Interpretating an utterance grammatically, and in the dialogue context, happens in parallel in the system. This way, the dialogue context can help constraining parsing.

Furthermore, also the situated context is (indirectly) involved. While interpreting an utterance linguistically, the system also attempts to connect utterance content (once it has been interpreted against the dialogue context) with the situated context [11]. Models of the situated context currently contain visual information, spatial organization, and situated action planning. The results of this process again feed back into the linguistic analysis.

What is important for the approach we present here in this paper is that we thus have access to the dialogue- and situated contexts, while the analysis of an utterance unfolds.

As we explain in the next section, this contextual information is exploited to build a *salience model* of the environment (§4.1). This salience model is subsequently used to compute *lexical activation* levels in our vocabulary (§4.2) and adapt the *word-class probabilities* of the language model (§4.3). Finally, once the new probabilities are estimated, they are added to the working memory and retrieved by the speech recognizer which incorporates them at runtime.

The above process is repeated after each detected change in the visual or linguistic context.

4 Approach

4.1 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.

In the future, other information sources could be added, for instance the possible presence of gestures [5], eye gaze tracking [18], entities in large-scale space [31], or the integration of a task model – as salience generally depends on intentionality [16].

4.1.1 Visual salience

Via the working memory, 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 (e.g. 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.

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 $1 \leq k \leq |\mathbf{E}_v|$.

4.1.2 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 discourse structure, or in terms of *linear recency* of mention [12]. Our implementation can theorically handle both types of linguistic salience, but, at the time of writing, only the linear recency is calculated.

To compute the linguistic salience, we extract a set $\mathbf{E}_l$ of potential referents from the discourse context model, and for each referent e_k we assign a salience value $s(e_k)$ equal to the linear distance between its last mention and the current position in the discourse model.

4.1.3 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:

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

where $I_A(x)$ is the indicator function of set A, and δ_v, δ_k are factors controlling the relative importance of each type of salience. They are determined empirically, subject to the following normalisation constraint: $\delta_v \sum_{e_k \in \mathbf{E}_v} s(e_k) + \delta_l \sum_{e_k \in \mathbf{E}_l} s(e_k) = |\mathbf{E}|$.

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

4.2 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, ‘close’, which is less situation specific.

The use of lexical activation networks is a key difference between our model and [21], which relies on a measure of “descriptive fitness” to modify the word probabilities. One key 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, ‘screen’ and ‘switch on/off’ would for instance be activated.

[2] built on top of the OpenCCG NLP library: http://openccg.sf.net

If the probability of specific words is increased, we need to renormalise 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 '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.

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

1. the set of physical objects the vision subsystem can recognise ;
2. the set of nouns specified in the lexicon as an 'object'.

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

4.3 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 salience model.

4.3.1 Corpus generation

We need a corpus to train a statistical language model adapted to our task domain, consisting of human-robot interactions related to a fixed visual scene. The visual scene usually includes a small set of objects (mugs, balls, boxes) which can be manipulated by the robot.

Unfortunately, no corpus of situated dialogue adapted to our task domain was available. Collecting in-domain data via Wizard of Oz (WOz) experiments is a very costly and time-consuming process, so we decided to follow the approach advocated in [29] instead and *generate* a class-based corpus from a domain-specific grammar.

Practically, we first collected a small set of WOz experiments, totalling about 800 utterances. This set is too small to be directly used as a training corpus, but sufficient to get an intuitive idea of the type of utterances in our domain. Based on it, we designed a task-specific context-free grammar able to cover most of the utterances. Weights were then automatically assigned to each grammar rule by parsing our initial corpus, leading to a small *stochastic context-free grammar*. As a last step, this grammar is randomly traversed a large number of times, which provides us the generated corpus.

4.3.2 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:

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \underbrace{P(\mathbf{O}|\mathbf{W})}_{\text{acoustic model}} \times \underbrace{P(\mathbf{W}|\mathbf{E})}_{\text{salience-driven language model}} \qquad (2)$$

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

$$P(w_1^n|\mathbf{E}) \simeq \prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2};\mathbf{E}) \qquad (3)$$

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:

$$P(w_i|w_{i-1}w_{i-2};\mathbf{E}) = \underbrace{P(w_i|c_i;\mathbf{E})}_{\text{word-class probability}} \times \underbrace{P(c_i|c_{i-1},c_{i-2})}_{\text{class transition probability}} \qquad (4)$$

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

$$P(w_i|c_i;\mathbf{E}) = \sum_{e_k \in \mathbf{E}} P(w_i|c_i;e_k) \times P(e_k) \qquad (5)$$

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

$$P(w_i|c_i;e_k) = \begin{cases} P(w_i|c_i) + \alpha_1 & \text{if } c_1 \\ P(w_i|c_i) - \alpha_2 & \text{if } \neg c_1 \wedge c_2 \\ P(w_i|c_i) & \text{else} \end{cases} \qquad (6)$$

where $c_1 \equiv w_i \in \mathsf{activatedWords}(e_k)$, and $c_2 \equiv w_i \in \mathsf{contextDependentWords}$. The optimum value of α_1 is determined using regression tests, while α_2 is computed relative to α_1 in order to keep the sum of all probabilities equal to 1:

$$\alpha_2 = \frac{|\mathsf{activatedWords}|}{|\mathsf{contextDependentWords}| - |\mathsf{activatedWords}|} \times \alpha_1 \qquad (7)$$

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

5 Evaluation

We evaluated our approach on a test suite of 250 spoken utterances recorded during WOz experiments. 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. The interactions could include descriptions, questions and commands.

Table 1 gives the experimental results. For space reasons, we focus on the WER of our model compared to the baseline (a class-based trigram model not using salience).

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

Table 1. Comparative results of recognition performance

5.1 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, the WER is indeed reduced by 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.

Figure 4. Sample visual scene including 3 objects: a box, a ball, and a chocolate bar

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. We estimated that 89 % of the recognition errors were due to function words. Moreover, our chosen test suite is constituted of "free speech" interactions, which often include lexical items or grammatical constructs outside the range of our language model.

6 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.

As future work we will examine the potential extension of our approach in two directions. First, we wish to take other information sources into account, particularly the integration of a task model, relying on data made available by the symbolic planner. And second, we want to go beyond speech recognition, and investigate the relevance of such salience model for the development of a robust understanding system for situated dialogue.

REFERENCES

[1] J. Baldridge and G.-J. M. Kruijff, 'Multi-modal combinatory categorial grammmar', in *Proceedings of EACL'03*, Budapest, Hungary, (2003).

[2] J. Baldridge and G.J.M. Kruijff, 'Coupling CCG and hybrid logic dependency semantics', in *Proc. ACL 2002*, pp. 319–326, Philadelphia, PA, (2002).

[3] M. Brenner, N. Hawes, J. Kelleher, and J. Wyatt, 'Mediating between qualitative and quantitative representations for task-orientated human-robot interaction', in *Proc. of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, (2007).

[4] T. Brick and M. Scheutz, 'Incremental natural language processing for HRI', in *Proceeding of the ACM/IEEE international conference on Human-Robot Interaction (HRI'07)*, pp. 263 – 270, (2007).

[5] J. Y. Chai and Sh. Qu, '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*, pp. 217–224, Vancouver, Canada, (October 2005). ACL.

[6] P. Gorniak and D. Roy, 'Grounded semantic composition for visual scenes', *Journal of Artificial Intelligence Research*, **21**, 429–470, (2004).

[7] P. Gorniak and D. Roy, 'Probabilistic grounding of situated speech using plan recognition and reference resolution', in *Proceedings of the Seventh International Conference on Multimodal Interfaces (ICMI 2005)*, (2005).

[8] P. Gorniak and D. Roy, 'Situated language understanding as filtering perceived affordances', *Cognitive Science*, **31**(2), 197–231, (2007).

[9] A. Gruenstein, C. Wang, and S. Seneff, 'Context-sensitive statistical language modeling', in *Proceedings of INTERSPEECH 2005*, pp. 17–20, (2005).

[10] N. Hawes, A. Sloman, J. Wyatt, M. Zillich, H. Jacobsson, G.-J. M. Kruijff, M. Brenner, G. Berginc, and D. Skocaj, 'Towards an integrated robot with multiple cognitive functions.', in *AAAI*, pp. 1548–1553. AAAI Press, (2007).

[11] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, '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).

[12] J. Kelleher, 'Integrating visual and linguistic salience for reference resolution', in *Proceedings of the 16th Irish conference on Artificial Intelligence and Cognitive Science (AICS-05)*, ed., Norman Creaney, Portstewart, Northern Ireland, (2005).

[13] P. Knoeferle and M.C. Crocker, 'The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking', *Cognitive Science*, (2006).

[14] G.J.M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N.A. Hawes, 'Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction', in *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, pp. 55–64, Aveiro, Portugal, (2007).

[15] G.J.M. Kruijff, H. Zender, P. Jensfelt, and H.I. Christensen, 'Situated dialogue and spatial organization: What, where... and why?', *International Journal of Advanced Robotic Systems, Special section on Human and Robot Interactive Communication*, **4**(2), (March 2007).

[16] F. Landragin, 'Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems', *Signal Processing*, **86**(12), 3578–3595, (2006).

[17] R. K. Moore, 'Spoken language processing: piecing together the puzzle', *Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing*, **49**, 418–435, (2007).

[18] Sh. Qu and J. Chai, '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*, pp. 284–291, (2007).

[19] D. Roy, 'Situation-aware spoken language processing', in *Royal Institute of Acoustics Workshop on Innovation in Speech Processing*, Stratford-upon-Avon, England, (2001).

[20] D. Roy, 'Grounding words in perception and action: Insights from computational models', *Trends in Cognitive Science*, **9**(8), 389–96, (2005).

[21] D. Roy and N. Mukherjee, 'Towards situated speech understanding: visual context priming of language models', *Computer Speech & Language*, **19**(2), 227–248, (April 2005).

[22] D. Roy and E. Reiter, 'Connecting language to the world', *Artificial Intelligence*, **167**(1-2), 1–12, (2005).

[23] T.P. Spexard, S. Li, B. Wrede, M. Hanheide, E.A. Topp, and H. Httenrauch, 'Interaction awareness for joint environment exploration', in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 546–551, (2007).

[24] M. Steedman, *The Syntactic Process*, The MIT Press, 2000.

[25] L. Steels, 'Semiotic dynamics for embodied agents', *IEEE Intelligent Systems*, **21**, 32–38, (2006).

[26] L. Steels, 'The symbol grounding problem has been solved. so what's next?', in *Symbols, embodiment and meaning*, eds., M. De Vega, G. Glennberg, and G. Graesser, Academic Press, New Haven, (2008).

[27] L. Steels and J-C. Baillie, 'Shared grounding of event descriptions by autonomous robots', *Robotics and Autonomous Systems*, **43**(2-3), 163–173, (2003).

[28] C. Theobalt, J. Bos, T. Chapman, A. Espinosa-Romero, M. Fraser, G. Hayes, E. Klein, T. Oka, and R. Reeve, 'Talking to godot: Dialogue with a mobile robot', in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002)*, pp. 1338–1343, (2002).

[29] K. Weilhammer, M. N. Stuttle, and S. Young, 'Bootstrapping language models for dialogue systems', in *Proceedings of INTERSPEECH 2006*, Pittsburgh, PA, (2006).

[30] T. Winograd, 'A process model of language understanding', in *Computer Models of Thought and Language*, eds., R.C. Schank and K.M. Colby, 152–186, Freeman, New York, NY, (1973).

[31] H. Zender and G.-J. M. Kruijff, 'Towards generating referring expressions in a mobile robot scenario', in *Language and Robots: Proceedings of the Symposium*, pp. 101–106, Aveiro, Portugal, (December 2007).