A Computational Model of Concept Generalisation in Cross-Modal Reference

Patrick McCrae[1**], Wolfgang Menzel[2], Maosong Sun[3]

1. CINACS Graduate Research Group, Department of Informatics, Hamburg University, 22527 Hamburg, Germany;
2. Natural Language Systems Group, Department of Informatics, Hamburg University, 22527 Hamburg, Germany;
3. Department of Computer Science, Tsinghua University, Beijing 100084, China

Abstract: Cross-modal interactions between visual understanding and linguistic processing substantially contribute to the remarkable robustness of human language processing. We argue that the formation of cross-modal referential links is a prerequisite for the occurrence of cross-modal interactions between vision and language. In this paper we examine a computational model for cross-modal reference formation with respect to its robustness against conceptual underspecification in the visual modality. This investigation is motivated by the fact that natural systems are well capable of establishing cross-modal reference between modalities with different degrees of conceptual specification. In the investigated model, conceptually underspecified context information continues to drive the syntactic disambiguation of verb-centred syntactic ambiguities as long as the visual context contains the situation arity information of the visual scene.

Key words: Vision-Language Interaction, Cross-Modal Reference, Syntactic Disambiguation

Introduction

Humans construe a largely consistent and uniform mental representation of the world surrounding them based on the sensory input received from multiple modalities. The process of information fusion ensures that different cross-modal perceptions integrate into a single, uniform percept. In cross-modal perception, congruent information from different input channels affirms the cross-modal percept while incongruent information in the input modalities gives rise to a perceptual conflict. These conflicts direct attention and perception such as to acquire further information in order to resolve the perceptual conflict.

In this paper we focus on the interaction between visual scene understanding and linguistic processing (henceforth: *vision-language interaction*). We examine a computational model for cross-modal reference formation between vision and language with respect to its robustness to conceptual underspecification. Specifically, we investigate the model's syntactic disambiguation behaviour under the influence of visual scene contexts that are conceptually underspecified with respect to the linguistic modality. The central question addressed here is whether visual scene information that is conceptually underspecified with regards to the information in the linguistic modality can still contribute enough additional information to direct the process syntactic disambiguation in the linguistic modality.

1 Cross-Modal Interactions in Natural and Artificial Systems

1.1 Human sentence comprehension

There is compelling empirical evidence to suggest that humans form cross-modal referential links during the earliest stages of linguistic processing. Cooper[1] observed a significant preference of subjects to fixate depictions of objects that were either directly named or

simply referred to in an auditorily presented linguistic stimulus. Fixations also increased when the objects were semantically related to the entities named in the linguistic stimulus.

Tanenhaus et al.[2] showed that different visual scene contexts give rise to different structural starting hypotheses when parsing the same syntactically ambiguous sentence. The observed eye fixations support the view that cross-modal reference is established very rapidly and in close temporal alignment with the unfolding linguistic stimulus.

Knöferle[3] conducted a number of experiments to confirm the formation of cross-modal reference between spoken language and visual scenes, not only at the level of participating entities but also for visually perceived events, actions and processes. We adopt the terminology of Barwise and Perry[4] and collectively refer to these verb-centred concepts as *situations*.
Knöferle also investigated the relative importance of visual scene information compared with lexical knowledge. Her findings support the view that readily available visual scene information has a stronger semantic influence on linguistic processing than the stereotypicality of situation participants.

Jackendoff's *Conceptual Semantics*[5,6] constitutes an overarching cognitive framework to account for how non-linguistic modalities and language interact with each other. Jackendoff's *Conceptual Structure Hypothesis* holds that all modalities, be they linguistic or non-linguistic in nature, interface at the level of *Conceptual Structure*, the single and uniform level of semantic representation[5]. Jackendoff argues that interactions between non-linguistic modalities and language proceed with semantic mediation in Conceptual Structure[5]. The influence of the non-linguistic modalities on syntax results from mappings between conceptual structures and syntactic representations based on correspondence rules in the syntax-semantics interface [6].
The notion of interfaces mapping between distinct representations can be re-formulated in terms of representations that mutually constrain each other. Visual scene information can be considered to constrain the set of possible interpretations of a given natural language utterance. This constraint-based view also underlies a number of computational implementation approaches to vision-language interaction that have been reported in the literature[7], [8], [9], [10], [11], [12].

2 Modelling Cross-Modal Reference

2.1 McCrae's computational model

McCrae proposes a cognitively-motivated computational model for the influence of visual scene information upon syntactic parsing[13],[14]. The model implements central aspects of Jackendoff's Conceptual Semantics. Semantic representations of visual scene context are employed to constrain the assignment of semantic dependencies in WCDG, a weighted-constraint dependency parser for German. Visual scene context is modelled in a single, unified semantic representation of linguistic and non-linguistic semantics – analogous to Conceptual Structure. Contextual constraints propagate into syntax via correspondence rules in the syntax-semantics interface.

C*ross-modal matching*[15], i.e., the mapping of entities from the linguistic modality to entities in the non-linguistic modalities is based on the conceptual compatibility of the concepts activated in the linguistic modality and those instantiated in the visual scene context. The model exploits contextually asserted thematic relations in the visual scene to modulate semantic attachment decisions in the linguistic analysis. If two words in the linguistic modality have been found to make cross-modal reference to entities in the visual scene, the linguistic dependencies between these words will be affected by the thematic relations asserted between the visual entities[16], [17], [18].

2.2 Syntactic disambiguation by visual context

Consider the following German sentence which contains a genitive-dative ambiguous subclause after the comma (ambiguous constituent italicised):

Er weiß, dass der Verehrer *der Schauspielerin* den Blumenstrauß schenkte. (1.1)

He knows that …

{ … the actress's admirer gave the bouquet. (1.2)
  … the admirer gave the actress the bouquet. (1.3) }

In the absence of extrasentential information to guide reading preferences this global structural ambiguity remains undecidable. To arrive at a structural analysis for globally ambiguous sentences, WCDG applies lin-

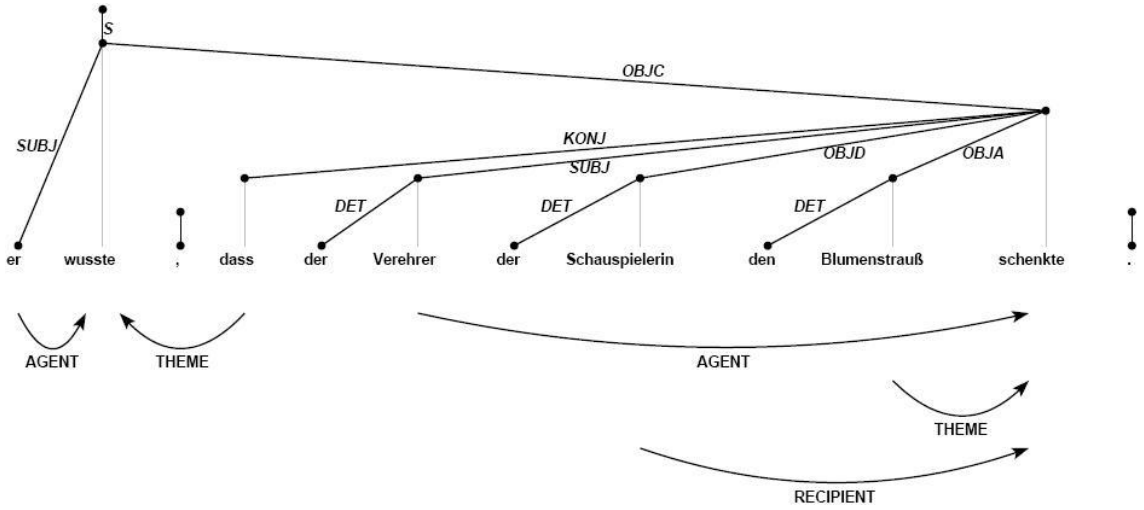


Fig. 1 WCDG's default analysis of genitive-dative ambiguitiy in the absence of a contextual bias.

guistic heuristics to favour one analysis over another. In the case of sentences with genitive-dative ambiguity, WCDG prefers the dative reading shown in Figure 1.
We can also drive the disambiguation by integrating a biasing visual scene context. A visual context in which the admirer of an actress is giving a bouquet will favour the binary reading. A context in which an admirer is giving an actress a bouquet will result in a preference for the ternary reading. In the computational model under investigation, these visual contexts are represented as shown in Figure 2.

The conceptual specification in these situation representations suggests that the visual modality provides precisely of the same degree of conceptual granularity as the linguistic modality. For a general interaction between visual scene context and linguistic processing, however, this is a rather unrealistic assumption.

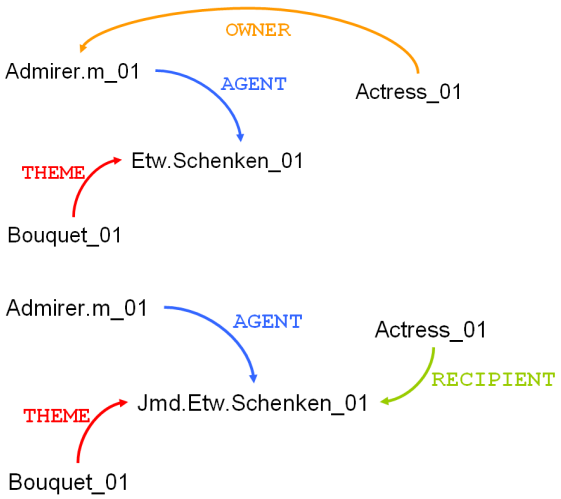


Fig. 2 Representations of the binary and the ternary situation.

In some cases, the visual modality will provide conceptually more specific information than the linguistic modality, while in other settings the opposite may be the case. As an example, consider the perceptual uncertainty resulting from suboptimal vision conditions as caused by insufficient lighting, scene occlusion or a large physical separation between the observer and the scene. Perceptual uncertainty yields conceptually underspecified visual percepts that instantiate concepts and thematic relations that are less specific.

This examination focuses on representations of visual percepts that contain instantiations of less specific entity and situation concepts. The concept generalisations needed to express this conceptual underspecification resulting from perceptual uncertainty can conveniently be obtained by exploiting ontological properties of the instantiated concepts: We simply replace the instantiations of the exact concepts by instantiations of superordinate concepts from the underlying conceptual hierarchy in the ontology's T-box.

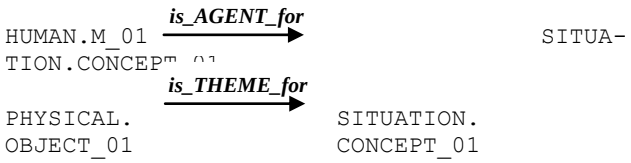
3 Experiments

3.1 Approach

Our experiments examine the effect of visual context information upon syntactic processing when the visual information is conceptually less specific than the given linguistic information. We parse 10 different sentences containing a global genitive-dative ambiguity. Each of these sentences has two readings analogous to 1.2 and 1.3 above. Just as for 1.1, we verified in advance that WCDG's default analysis in the absence of a contextual bias is the ternary (dative) reading.

We then integrate binary context models that instantiate concepts that are higher up in the T -Box's conceptual hierarchy – and hence are more general – than the concepts activated in the linguistic modality. Concept generalisations have been selected based on the following guidelines: Concepts denoting concrete entities are generalised to a visually perceivable superclass, e.g., ADMIRER is generalised to HUMAN.M and ACTRESS to HUMAN.F. Abstract concepts such as MOOD or ADDRESS are represented by ABSTRACT, their next higher superclass in the ontology. Concepts denoting inanimate entities are generalised to the superclass PHYSICAL.OBJECT. Verb-specific concepts are abstracted to a level at which the verb-specific properties regarding lexicalisation and precise thematic role subcategorisation are lost.

Since these generalised binary context models are meant to represent visual percepts under uncertainty, the question arises to what extent the modelled information can really be attained from a visual scene. A critical analysis of the context model generalisation for 1.2 motivates an additional modification to the generalised context models: The *is_OWNER_for* relation as such is visually not perceivable. We may want to include it, as argued for by McCrae[18], to reflect world knowledge of known entities. However, the perceptual uncertainty we are attempting to model in this case prevents entity recognition and hence requires the elimination of the *is_OWNER_for* relation from our context models. A context representation of

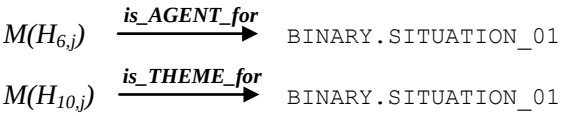


is therefore cognitively more plausible for 1.1 than the inclusion of the additional assertion of an *is_OWNER_for* relation. If the visual information is so uncertain that is does not permit the identification of HUMAN.M_01 as an ADMIRER_01, then the world-knowledge-based association with ACTRESS_01 or its generalised instance HUMAN.F_01 via an *is_OWNER_for* relation also cannot occur. The reduced context representations are cognitively more plausible because they only contain information that can actually be extracted from a visual scene under perceptual uncertainty.

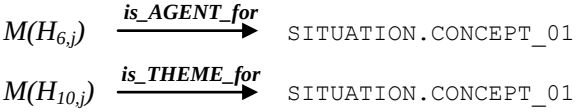
We now investigate whether the information provided in this generalised, reduced context model is still sufficient to constrain the parser's linguistic analysis to the context compliant non-default binary analysis. We wish to gain insight into how strongly the situation representation of a visual scene can be generalised in order to still afford the non-default linguistic analysis. In Experiment 1 we study if visual context models centred around an instance of BINARY.SITUATION are still restrictive enough to drive the syntactic modulations required for the non-default binary analysis. This situation concept has lost all verb-specific information except for the situation arity, i.e., the information about how many entities participate in the situation. In Experiment 2 we further generalise the central situation concept and instantiate SITUATION.CONCEPT, the most general situation concept. The resulting context representation is so general that *all* verb-specific information of the observed visual scene – including situation arity – is lost.

3.2 Setup

We parse 10 randomly selected globally ambiguous sentences of German taken from a psycholinguistic study[19]. Like 1.1, all of these contain an unambiguous introductory main clause followed by a globally ambiguous subclause with genitive-dative ambiguity. Normalisation of the main clauses yielded sentences of the generic pattern '*Er wusste, dass* ***A B C V***', where *A* is the subject, *B* the genitive-dative ambiguous constituent, *C* the accusative (direct) object and *V* the full verb. Reduced binary context models were prepared manually for all of these sentences. Let $M(H_{i,j})$ denote the cross-modal match of the *j*-th homonym in slot *i* of the input sentence. The reduced binary context models for Experiment 1 take the form



Analogously, for Experiment 2:



The sentences are parsed under soft integration, i.e., the parser may assign dependencies that are incompatible with the integrated context model. Doing so will, however, incur a score penalty for the parse.

4 Results

The parse trees obtained in Experiment 1 all comply with the structural scheme in Figure 3. Structurally, the trees for integration of the generalised contexts are found to be identical with those obtained under soft

integration of the conceptually specific context models.

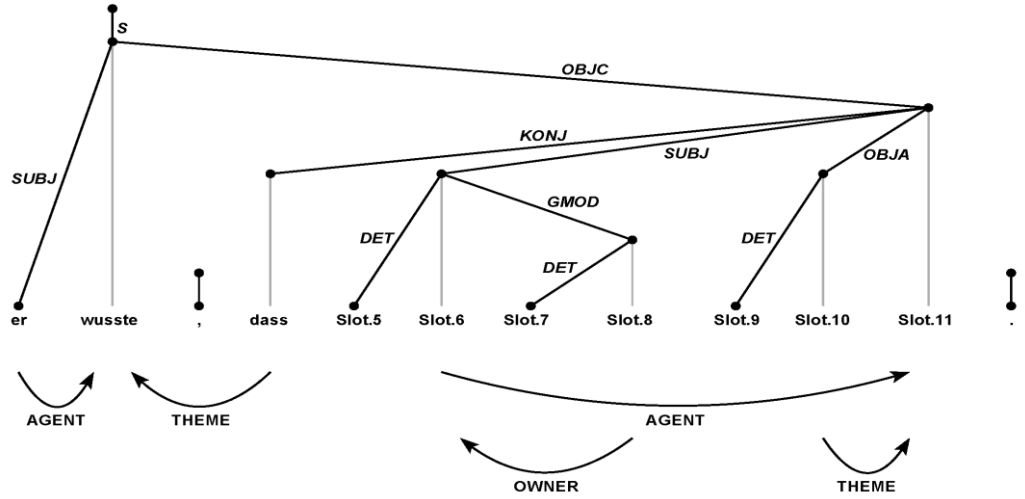


Fig. 3 Generic structural pattern for the binary analysis.

The reduction of the context models was consequently found to have no adverse effect on the context's ability to achieve disambiguation in the form of the non-default binary analysis.

The analyses obtained for Experiment 2 exhibit a pattern, the cause for which will be discussed in the following section. The majority of the parse trees follow the structural scheme in Figure 4, while three of the sentences follow the structural scheme in Figure 3. Using WCDG's capability to score manually modified parse trees, we were able to exclude search errors as a possible cause for the difference in analyses: for these three sentences the binary analysis does indeed receive a better score than the ternary analysis.

5 Discussion

The results of Experiment 1 clearly show that the integration of context models centred around an instance of `BINARY.SITUATION` is successful. This experiment models the influence of visual scene context upon linguistic processing when the situation in which participants interact with each other cannot be identified precisely. The only situation information visual context provides in these cases is the arity of the interaction between the observed entities. With the integration of a binary visual context the model effects the dismissal of all ternary verb forms as possible readings. Note that the investigated genitive-dative ambiguity has an effect on verb valence: the `GMOD/OWNER` reading requires the binary verb form while the `OBJD/RECIPIENT` reading needs the ternary verb form. We consequently expect visual context to lose its constraining effect upon the resolution of genitive-dative ambiguity when the instantiated concepts become so general that their situation arity information is lost. They will then fail to restrict the selection of homonyms with the appropriate valence in the parser.

Increasing the generality of concepts instantiated in visual context typically has two effects: both the number of cross-modal matches per homonym and the number of homonyms receiving a cross-modal match increase. The less specific a modelled visual percept is conceptually, the less constraining its effect upon linguistic processing will be.

As expected, the integration of the two-entity contexts centred around an instance of `SITUATION.CONCEPT` fails to induce the binary analysis consistently. Most of the structures afforded are expectation compliant and follow the structural paradigm of the ternary situation analysis, i.e., they comply with the linguistic default preferences.

If the hypothesis is correct that contexts instantiating `SITUATION.CONCEPT` cannot drive the binary analysis, why then do not *all* sentences in Experiment 2 afford the ternary analysis? The reason for this becomes apparent when we consider the integration constraints that, according to WCDG, are violated by each of the solution structures. All sentences affording the default analysis violate three integration constraints, namely for the `AGENT`, `THEME` and `RECIPIENT` assignment. The other three sentences that have been analysed contrary to expectation only violate the integration constraint for the `THEME` dependencies.

The reason for this is as follows: In contrast to the expectation-compliant sentences, the other three sen-

tences all integrate a context model that asserts the entity HUMAN.M_01 as an AGENT for an instance of SITUATION.CONCEPT. Therefore, the AGENT edges between Slot.1 and Slot.2 as well as between Slot.6

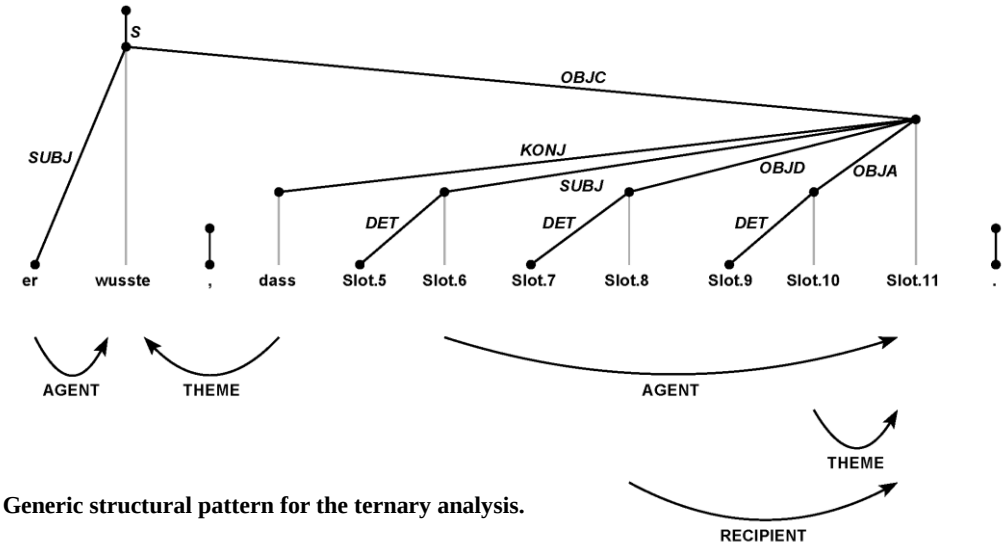


Fig. 4 Generic structural pattern for the ternary analysis.

and Slot.11 are contextually compatible and do not cause a constraint violation.

Moreover, the parser resolves the contextual constraints on the genitive-dative ambiguity in favour of the binary analysis to reduce the number of contextual constraint violations overall. Since we are integrating a reduced binary context model that does not contain an *is_OWNER_for* assertion anymore, the OWNER dependency can be assigned from Slot.8 without the violation of an integration constraint.

The integration constraint violation on for the *is_THEME_for* dependency remains because Slot.2 still matches cross-modally with SITUATION.CONCEPT_01. The latter concept instance, however, already has a *is_THEME_for* assertion from another slot, namely Slot.10, such that the model vetoes all other incoming *is_THEME_for* dependencies.

A final comment is owed to the cognitive plausibility of visual contexts centring around instances of SITUATION.CONCEPT. Effectively, these are visual contexts in which the information contained is so general that neither the nature of the interaction between the observed entities nor the arity of the interaction are known. In our view it is highly questionable whether the instantiation of such concepts can serve a cognitive purpose – and hence whether such percepts realistically arise at all. We rephrase this doubt as the question of whether SITUATION.CONCEPT is encoded in the human cognitive system at all. A substantial amount of further investigation in the area of cognitive psychology and cognitive science will be needed to answer this question conclusively.

For the investigation of the computational model, we conclude that the generalisation of the central situation concept to a degree at which situation arity information is lost, results in the breakdown of its power to effect the systematic disambiguation in verb-related syntactic ambiguities such as genitive-dative ambiguity.

6 Conclusions

The two experiments reported here addressed the question how strongly we can generalise the concepts instantiated in a visual context representation in order to still achieve a disambiguating cross-modal influence upon linguistic processing. The degree of permissible concept generalisation depends on the type of syntactic ambiguity in the input sentence as well as on the concept properties modelled in the underlying ontology's T-Box. We have seen that for a syntactic ambiguity type that affects verb valence reliable syntactic disambiguation requires the availability of situation arity information from visual context. The resolution of syntactic ambiguities that do not affect verb valence, such as PP-attachment, can be achieved with visual contexts that are conceptually specific enough to yield different attachment predictions for the constituents in question. A visual context that is so general that it effects the same predictions for all words in the sentence, e.g., a context model instantiating only instances of THING or TOP, will lose all of its potential to constrain linguistic analysis in the given model. We have furthermore expressed substantial doubt as to whether

the visual modality in humans will give rise to mental representations that instantiate extremely general concepts such as `SITUATION.CONCEPT` or `THING`.

7 Future Work

While this paper has investigated the effect of concept generalisations on context integration in a computational model for vision-language interaction, the generalisation of the thematic relations between the contextual entities has been left untouched. It may well be the case that the perception of a visual scene also results in the assignment of more general – and hence more ambiguous – thematic relations than those considered in this work. We encourage the exploration of this field in future research endeavours, both from the perspective of cognitive science and as a potential extension to the modelling capabilities of the investigated framework.

Acknowledgements

We gratefully wish to acknowledge the financial support of this work by German Research Foundation grant GRK 1247/1.

References

[1] Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. a new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology, 6:84–107.

[2] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. SCIENCE, 268:1632–1634.

[3] Knöferle, P. S. (2005). The Role of Visual Scenes in Spoken Language Comprehension: Evidence from Eye-Tracking. PhD thesis, Universität des Saarlandes Saarbrücken, Germany.

[4] Barwise J, Perry J. Situations and Attitudes. Cambridge, MA: MIT Press, 1983.

[5] Jackendoff, R. (1983). Semantics and Cognition. Cambridge, MA: MIT Press.

[6] Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F., editors, Language and Space, chapter 1, pages 1–30. Cambridge, MA: MIT Press.

[7] Brown M K, Buntschuh B M, Wilpon J G. Sam: A perceptive spoken language understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 1992, **22**: 1390–1402.

[8] Srihari, R. K. and Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In: Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), Seattle, pages 793–798.

[9] Socher, G. (1997). Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding. PhD thesis, Technical Faculty, University of Bielefeld, Germany.

[10] Socher, G., Sagerer, G., Kummert, F., and Fuhr, T. (1996). Talking about 3d scenes: Integration of image and speech understanding in a hybrid distributed system. In Proceedings of the International Conference on Image Processing (ICIP-96), Lausanne, page 18A2.

[11] Socher, G., Sagerer, G., and Perona, P. (2000). Bayesian reasoning on qualitative descriptions from images and speech. Image And Vision Computing, 18(2):155–172.

[12] Brick T, Scheutz M. Incremental natural language processing for hri. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction. Washington DC, 2007: 263–270.

[13] McCrae, P. (2007). Integrating cross-modal context for PP attachment disambiguation. In: Proceedings of the 3rd International Conference on Natural Computation (ICNC 2007, Haikou, China), volume 3, pages 292–296. Los Alamitos, CA: IEEE.

[14] McCrae, P. and Menzel, W. (2007). Towards a system architecture for integrating cross-modal context in syntactic disambiguation. In: Sharp, B. and Zock, M., editors, Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007, Funchal, Portugal), pages 228–237. INSTICC Press.

[15] Bushnell, E. W. (1994). The development of intersensory perception: comparative perspectives, chapter A Dual-Processing Approach to Cross-Modal Matching: Implications for Development, pages 19–38. New Jersey: Lawrence Erlbaum Associates.

[16] McCrae, P. (2009a). How reasoning achieves context integration into syntax parsing. In: Proceedings of the 2009 Workshop on Intelligent Linguistic Technologies (ILINTEC 09) as part of the 2009 International Conference on Artificial Intelligence (ICAI 09, Las Vegas, USA), pages 455–461. CSREA Press.

[17] McCrae, P. (2009b). A model for the cross-modal influence of visual context upon language processing. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 09, Borovets, Bulgaria), pages 230–235. Shoumen: INCOMA.

[18] McCrae, P. (2010). A Computational Model for the Influence of Cross-Modal Context upon Syntactic Parsing. PhD thesis, Department of Informatics, University of Hamburg, Germany.

[19] van Kampen, A. (2001). Syntaktische und semantische Verarbeitungsprozesse bei der Analyse strukturell mehrdeutiger Verbfinalsätze im Deutschen: Eine empirische Untersuchung. PhD thesis, Freie Universität Berlin, Germany.