Connectionist Modeling of Situated Language Processing: Language and Meaning Acquisition from an Embodiment Perspective

Helmut Weldle (helmut.weldle@misc.uni-freiburg.de), Lars Konieczny (lars@cognition.uni-freiburg.de), Daniel Müller (daniel@cognition.uni-freiburg.de), Sascha Wolfer (sascha@cognition.uni-freiburg.de), Peter Baumann (peter.baumann@cognition.uni-freiburg.de)

Center for Cognitive Science, University of Freiburg, Friedrichstr. 50
D-79098 Freiburg i. Br., Germany

Abstract

Recent connectionist models and theories of embodied cognition offer new perspectives on language comprehension. We review the latest accounts on the issue and present an SRN-based model, which incorporates ideas of embodiment theories and avoids (1) vast architectural complexity, (2) explicit structured semantic input, and (3) separated training regimens for processing components.

Keywords: language acquisition, comprehension, production; sentence processing; language-vision integration; visual attention; embodied cognition; connectionist modeling; SRNs.

Introduction

'Gavagai!' If we heard a native speaking a foreign language utter this word upon seeing a rabbit, we would be faced with the problem Quine described in *Ontological Relativity* (1968): How do we know what exactly an utterance refers to in an infinitely rich set of objects, events and relations our environment provides? But this problem appears almost trivial compared to a human child confronted with the task to acquire its mother's language. Several sub-tasks have to be solved simultaneously to achieve this grounding of speech to referential meaning: there is the problem of a highly complex world rich in details, happenings and relations. There is the problem of a continuous stream of words. There is the large problem of relating the one to the other. And there is the problem that there is no previously given language to help finding this relation.

In other words, the task is to bind a holistic situation to a sequential series of related linguistic expressions. This affords to integrate representations of language and the outside world, both represented in distinctive forms, following completely different rules and depending on different hierarchical and causal relations. A central aspect of models of language comprehension and acquisition is how they account for these questions. In connectionist models, language interpretation and integration of situational context is based on mechanisms of association and self-organization.

Theories in embodied cognition research offer an account for assignment of linguistic structures to constructions of coherent semantic interpretations. Language comprehension is considered to be a simulation of perceptual experiences of the hearer, and the linguistic structure serves as an instruction for the correct construction of the situation. Due to its analogical nature this could be a guideline for subsymbolic accounts for grounding language comprehension.

Connectionist models of language comprehension

Several connectionist architectures deal with the task of language comprehension and integration of language and events, proposing different realizations of semantic representation and implementations of the integration process.

Rohde (2002) introduced the Connectionist Sentence Comprehension and Production Model, an architecture based on extended simple recurrent networks (SRNs, Elman, 1990) which is capable of comprehending and producing complex sentences, covering a wide range of well-known empirical phenomena. The model clearly focuses on scalability, however possibly at the expense of explanatory power and psychological plausibility. Especially relevant for our issue is the realization of the semantic component: Rohde's model is trained with explicit propositional representations prior to the corresponding target sentence. This greatly assists the network, leaving no way to tell whether it simply succeeds because the semantic representation provided all crucial information explicitly. Learning of the propositions is achieved through a query mechanism, inquiring each of its parts, a process questionable in its cognitive adequacy.

The problem of explicit information holds similarly for the Incremental Nonmonotonic Self-organization of Meaning Network (Mayberry, 2003). The semantic representations used in the model are based on Minimal Recursion Semantics (Copestake et al., 2005), which makes them very powerful and complex information carriers. The model is capable of parsing natural language corpora, an impressive achievement, reached at the expense of a highly complex, opaque architecture and pre-fabricated semantic content. In a more recent study Mayberry, Crocker and Knoeferle (2005) introduced the Coordinated Interplay Account Network which integrates a scene representation with the incremental input of a sentence description, enabling adaptive use of context information. Since the presented scenes are externally segmented into agent, action and patient, the major part of semantic interpretation is provided to the model explicitly.

While these models certainly achieved good results concerning their aims, they show several shortcomings making them unsuitable for our approach. Firstly, extensive use of different layers and components makes it impossible to deduce responsible structures and working mechanisms from internal states of the model, prohibiting analysis of ongoing processes. Secondly, reliance on explicit, extremely powerful semantic representations that do most of the work of semantic processing prevents clear assignment of performance properties to inherent connectionist mechanisms. Thirdly, the use of separated training for the components (e.g., pre-training of the semantic layer before coupling it with sequential linguistic input) reduces integration to an interface between independent modules. This contradicts the idea of grounded language acquisition, inhibiting examination of mutual and synergetic effects between syntactic and semantic components.

The Distributed Situation Space model (Frank et al., 2007) does not use explicit propositions. It uses Self Organizing Maps (Kohonen, 1995) to represent simple and combined events in a microworld and maps descriptive sentences on corresponding situation vectors using SRNs. The model implements the idea of non-propositional semantic representations, preserving analogy of internal representations to external states on the dimension of transition and combination probabilities of events. Frank, Haselager and van Rooij (2009) explored the capability of this model to capture semantic systematicity beyond simply implementing symbolic computation. They concluded that connectionist systematicity emerges from interaction with the environment, reflecting the observed and derived structural correlations. Considered from an embodied cognition perspective, the model preserves analogy in its internal representations. But it still waives central qualities that need to be explored: it reduces semantic content to co-occurrences of events and does not aim at modality at all, thereby leaving out inferences on behalf of event-internal relations.

Embodiment theories of language comprehension

Embodied cognition posits that the structure of embedded systems emerges as a consequence of interaction with the environment. This leads to an alternative perspective on cognitive processes and conceptualizations, which is highly compatible with recent connectionist and emergentist assumptions and enables the development of proposition-free comprehension systems.

With the Perceptual Symbol Systems framework, Barsalou (1999) emphasizes cognition to be grounded in perception, operating on modal and analogue symbols. These are derived directly as neural substrates of activations corresponding to perceptions of the external world and share the same functional brain areas. This is claimed to obviate the grounding and transduction problem. Language comprehension is seen as a mental simulation process of perceptual states of neural activation, triggered by linguistic input. Joyce et al. (2003) proposed connectionism as a suitable framework for closing some explanatory gaps concerning the question, how such a system could actually be implemented. Based on an SRN-model of perceptual symbol formation they drew further specifications of the required mechanisms and how they come to work.

Zwaan (2004) introduces an explicit framework for embodied language comprehension that integrates several empirical findings. In his Immersed Experiencer Framework, situational entities correspond to activated functional neural webs instantiated by lexical items. These webs become integrated to construals by means of constraint-satisfaction mechanisms, representing events corresponding to clauses. The approach claims to replace propositional representations, which are stated to be merely illustrative shorthand.

Modeling situated language comprehension

Our model directly addresses the discussed issues by imposing restrictions on the architecture, the nature of representations and tasks. Basically we assume that meaning is not an inherent feature of language, but must be assigned by grounded language acquisition: the meaning of a linguistic expression is the activated mental simulation of the corresponding event. Interpretation processes are guided by mechanisms of constraint-satisfaction, naturally inherent in artificial neural networks. Our focus lies on the integration of sequential linguistic and static situational information. The network is trained to achieve the simultaneous completion of different tasks: prediction of the sequential succession of linguistic units and recognition and classification of visual patterns related to the linguistic input. The tasks pose different requirements: extraction of sequential structures and their probabilities as well as extraction and generalization of diverse static patterns. Our aim is to explore the ability of the model to integrate these tasks, the mechanisms to map the contents of the differing information systems, and the usage of corresponding information of the respective system as an additional source of constraints.

Architecture and flow of information

We tried to avoid the inflationary use of hidden layers and different modules to keep functional assignments transparent and interrelated effects of the components analyzable. The base architecture is an SRN, with different input and output layers for linguistic and visual/situational information processing (Figure 1).

The syntax component[1] gets linguistic input (at Input I) in a sequential manner – sentences word by word – and is trained to predict the next input (at Output I), thereby extracting word classes, word transition probabilities and exploring structural relations in the input. Activation is forwarded from Input I over the Integration Layer to Output I. Learning of sequential structures is enabled through a context layer, which provides information about previous states of activation by copying the hidden layer and returning it at the next cycle.

The situational component propagates compressed simple static situations and is trained to reproduce its crucial information. The actual representations of the situational component are provided by the visual input and teaching examples: Input II presents static situations with simple

[1] The labels syntax and semantics are used as abbreviations for sequential linguistic vs. situational semantic component.

objects on a two-dimensional grid, constituting a retina-like interface for the model. Activation is forwarded to the Integration Layer through an additional hidden layer (Encoder: Semantics) encoding a distributed representation of Input II and reaches Output II via a second additional hidden layer (Decoder: Semantics). Output II provides a prototype version of the scene, forcing the network to extract crucial information about space and objects.

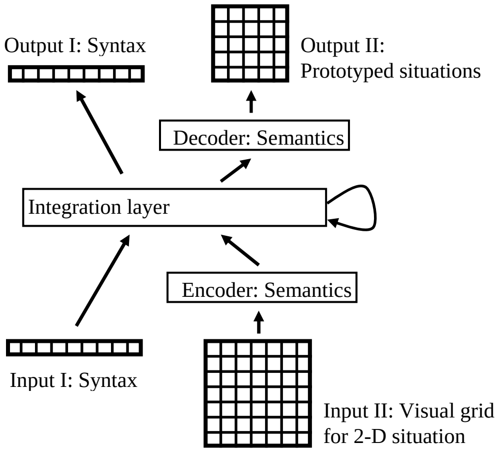

Figure 1: Extended SRN-architecture for integration of linguistic and situational information.

Both types of information – static situation and corresponding sequential linguistic input – are presented simultaneously. The situation is maintained as long as linguistic processing is in progress. The information from both components is integrated in the Integration Layer – allowing generalization over co-occurrences of incremental linguistic and static situational input. The network is forced to find common representations to solve the differing tasks of both components. In contrast to other connectionist models covering sequential tasks, we did not use a separated training regimen for the syntax and semantics components, but trained the complete network in a holistic fashion on the different information sources.

Semantics and situational representation

Following the idea of embodied language comprehension theories, conceptualization is assumed to be modal and analogue, encoded as schematic abstractions of events and objects in the environment. Meaning is assigned to linguistic units on different levels by mapping the units onto the internal structure that simulates the corresponding object, action or event structure. We can distinguish between a non- or pre-linguistic representation that conceptualizes the perceived world independently of linguistic labeling and a linguistic semantics that is constituted by assignment processes, corresponding to the relation between prototype theory and prototype semantics (Rosch, 1978). So, semantics is seen as a process rather than a state, led by constraint-satisfaction mechanisms, which dynamically adjust categories based on new experiences. We assume an intertwined development of positively interacting systems, as proposed in syntactic and semantic bootstrapping theories.

Visual patterns. To generate visual patterns, we used four discrete distinctive objects, named minus (-), pipe (|), slash (/) and backslash (\), distributed freely on a two-dimensional grid. Up to three of these objects were placed in different locations on that grid. Our situations differ in the number of involved objects, the identity of the selected objects, the exact location of these objects, and, as a consequence, in the spatial relation between those objects. The left panel of Figure 3 in the results section provides an example of an input situation depicting a backslash positioned roughly above the minus and slightly left above the pipe (this is obviously just one of several equally possible descriptions).

Concerning the retina-like implementation: our intention is not to provide a cognitively plausible model of visual processing, as for example realized by Coventry et al. (2005) in the Functional Geometric Framework. The retinal grid merely provides an intuitive and simplistic presentation format and offers several advantages: orientation towards modal features, analogy on the spatial dimension on the situational level and sparsely structured, non-explicitly encoded access to information about the situations.

The target grids contain prototype versions of corresponding input situations, reduced in several ways: they represent only two objects of an arbitrary number of initially presented objects, reflecting attentional focus on selected entities. It reduces the spatial expansion and idealizes the relative positions of considered objects to a prototypical spatial relation. Again using Figure 3 as illustrative example, the selected prototype for the input situation is the backslash positioned directly above the minus, blinding out the pipe. Mapping onto prototypes forces the network to instantiate self-organized internal representations of the situations that are selective and schematic in nature, extracting relevant information. This enables the model to distinguish the objects and to develop the concept of relative spatial positions, a presupposition for the mapping of corresponding linguistic input. Dominey (2003) demonstrated such a purely associative mechanism to be sufficient to inductively acquire productive grammatical constructions.

Table 1: Syntactic inventory for situation description.

object-A be-located deictic-particle eos deictic-particle be-located object-A eos
object-A be-located location-relative object-B eos object-B be-located inverse-location-relative object-A eos
location-relative object-A be-located object-B eos inverse-location-relative object-B be-located object-A eos

We used a very limited microlanguage for situation description (Table 1). The linguistic input consists of sentences presented sequentially word by word. The vocabulary contains lexical units for the distinguishable objects, for relative positions of the objects in space, a state verb, a deictic particle and an end-of-sentence marker. It is encoded in a localist fashion, each active unit representing one word. Visual situations are complex in that they entail several possible spatial relations between several objects. By expressing one of these relations between exactly two objects, the network is forced to direct its focus of attention.

Hypotheses on network performance

We expect language to have a positive effect on the discrimination and categorization of the components establishing the corresponding situations. Vice versa, we expect beneficial effects of the visual input on word prediction, based on the assumption that the network makes use of mutual informative cues. This is not as obvious as it may appear, since one could just as well assume that the tasks interfere with each other due to their different representation formats and the narrow resource of representational space.

In the test phase the network should be able (1) to derive the situational prototype without relying on the information of the situational input by exclusively using the sequential linguistic input and (2) to derive a coherent and correct sequential linguistic output, exclusively using the information of the static situation input. Furthermore the network is expected (3) to show systematic performance when coping with ambiguous cases occurring several times. When tested on the complete information including vision and language, we expect the network (4) to use the integrated crossover information to draw inferences on the correct output and to rule out irrelevant possibilities.

Materials and simulations

We ran the simulations on LENS (Rohde, 1999). Language input and output layers consisted of 16 units each, vision input had a grid size of 9 x 9 (= 81) units, prototyped vision output had a grid of 7 x 7 (= 49) units. Hidden layers consisted of 40 units each. The presented visual scenes showed up to three out of four different objects (minus, pipe, slash, backslash), forming complex spatial relations. The training sets contained 60% randomly chosen situations of the 130708 possible constellations of objects.

The training regimen consisted of three different stimuli settings. (1) *trainComplete* provides visual and linguistic stimuli in parallel, (2) *trainProduction* provides only visual stimuli to force the network to perform language production for observed scenes, (3) *trainComprehension* provides only linguistic stimuli forcing the language comprehension process. The target information always included both visual scene and corresponding linguistic description. The three training conditions were presented in an alternating manner.

The networks were trained for 10 epochs, but already showed very good performance on early epochs. We used the backpropagation-through-time algorithm, applying momentum 0.3, an initial weight range of 0.3 and a learning rate of 0.2, which was incrementally decreased by a factor of 0.02 per epoch.[2] The language output layer received a softmax function to enforce output activation complying with an interpretation of word probability. The test set used a sample of scene constellations excluded during the training phase. It was constructed corresponding to the training stimuli, providing full or single-sided reduced information: *testComplete*, *testProduction* and *testComprehension*.

Results and discussion

The proposed architecture has the potential to simulate and predict behavior relevant to research areas ranging from language acquisition to visual attention. What we present here are preliminary results of network performance using a subset of the possible training and test variations. The results are reported in three sections, treating (1) vision/comprehension, (2) word prediction and (3) language production, including relevant aspects concerning their interaction.

Vision/Comprehension In the *testComplete* condition with two objects in the input the model manages to produce the correct vision prototype at time step 1 already, without showing wrong objects. Over successive time steps, already clear objects get activated even stronger receiving additional support from the linguistic input. In ambiguous visual situations the vision output provides a preferred reading (due to slightly different frequencies in the training set). If the interpretation is falsified by the linguistic input, it adjusts the visual image to the linguistically referred position immediately.

In the three objects condition, the vision output first provides a diffuse but by no means arbitrary activation pattern: it contains all possible objects in their possible relations, in most cases with preferences for one constellation. During the time course of the incoming sentence, affirmation of the selected constellation by the linguistic description guides the vision output to adjust the pattern to the predetermined state, using at each step the available information at hand. New information reducing the possible constellation leads directly to a correction. At the last step only the requested pattern is produced, leaving no deviant activation (compare Figure 2). Importantly, at the final steps in both conditions the vision output contains only the activation of the correct predetermined constellation.

For the two-objects condition, this adjustment process could best be described in terms of disambiguation and discrete categorization. The process for the three-objects condition is rather a shift of attentional focus. Corresponding to the linguistic information, irrelevant aspects of the scene are shifted out of the focus of interpretation, while the relevant information receives highlighting by exclusive activation (Figure 3). This high context sensitivity emerges from implicit constraint-satisfaction mechanisms of the network, inherently establishing a frame-of-attention mechanism not built into the system artificially.

[2] We ran several sets of networks, varying learning rate (0.05 to 0.2), momentum (0.0 to 0.6) and initial weight range (0.1 to 1.0). Results were largely robust for most combinations of parameters, showing slight performance loss for some combinations, but preserving the qualitative systematicity of performance.

In the *testComprehension* condition we can observe the same behavior of direct correspondence between provided linguistic information and constructed scene prediction. Slightly different from the *testComplete* set, it produced even clearer constellation predictions, since input is less distorted by competing relations. The comprehension process succeeds incrementally, using all new information to construct a scene prediction as complete as possible. As soon as all relevant information is provided by the linguistic input, a stable visual scene prediction is constructed, containing only correct activations. This proves that the comprehension mechanism works independently from provided visual information and is able to construct complete visual scenes from linguistic descriptions, when no visual information is provided.

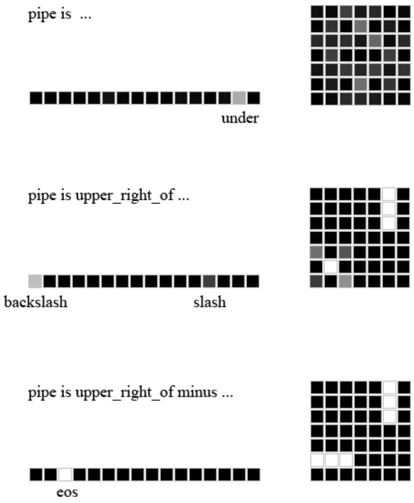

Figure 2: Comprehension and inference effect.

During the incremental construction of the situation model one can observe different strategies for using given and inferring missing information. As soon as a spatial relation is referred to, the vision component displays placeholders for the respective positions. In some cases, these take the shape of some object, indicating an expected default. If an object is already given, it is positioned correctly, leaving the second underspecified. With no spatial relation given, an introduced object is instantiated with weak and slightly obscured activation on multiple positions.

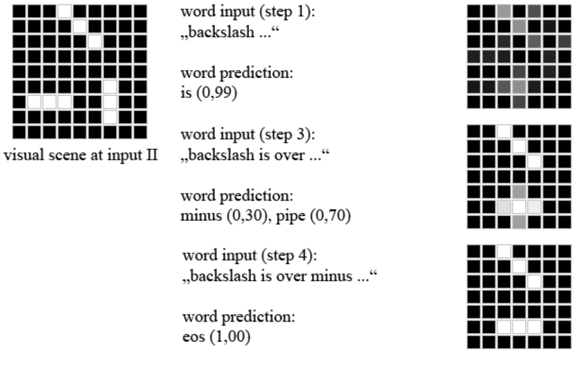

Figure 3: Comprehension and attentional focus effect.

Word prediction. Like the vision output the linguistic output is highly accurate in the *testComplete* condition with two and three objects. At time step 1 the network prefers to predict the spatial relations. It states only possible relations according to the visual input. The same holds for the following time steps. During object prediction, the network predicts only objects observed at the visual retina. At functional syntactic positions (verb, end-of-sentence marker), the network produces optimal activation.

Moreover, in successive time steps, it predicts only those objects and relations that can occur corresponding to the descriptive sequence given before (e.g., after the object location 'right-of', only objects are linguistically predicted, that can occur in a 'right-of' constellation). So it achieves far more than POS-tagging or grammatical probability matching (as criticized by Steedman, 1999). Its predictions are sensitive to word transition probabilities and grammatical category, and at the same time sensitive to possible descriptions of the scenes as provided by the visual context. We interpret this as visual priming, determining the linguistic performance by pre-activation of the relevant lexical items. In the condition *testComprehension*, the word prediction component could not rely on additional visual information. Therefore it activated all possible objects and relations at the respective syntactic positions, which are the only constraints provided by the preceding linguistic information.

Language production. For the *testProduction* set we reconfigured the trained network with a copy-connection from the language output layer (Output I) to the language input layer (Input I) and equipped the output layer with a winner-take-all functionality.[3] We have just begun analyzing production data, so the results are somewhat preliminary. The network produces mainly correct and complete sentences. These sequences contained only objects and relations that were given in the visual input, however not always expressing the correct constellation of objects. The network always instantiated linguistic starting points driven in correspondence to the distinctiveness of the visual input (e.g., clarity of spatial relation). While the network did rarely produce wrong sentences, it sometimes produced sentences that did not accurately correspond to the visual scene. Sometimes, objects were repeated (slash is left-of slash), or objects appeared in reverse order. The majority of sentences however expressed the correct constellation of objects.

The successful integration of information and interaction between the two components is reflected in the strong correspondence between language and vision: even when the network assumes a wrong constellation (e.g., for ambiguous cases and for three objects with initially wrong preferred constellations), this is predicted consistently on the vision and language output. Moreover, activation strengths for the objects and locations in the visual output and their respective word nodes correspond to each other (Figure 2, second

[3] This feedback loop triggers sequential routines to control the production process, providing the actual output as input stimulus on the next time step (e.g., Rohde, 2002).

pattern). Furthermore, new linguistic information can lead to a reconstruction of the visual prototype preferred so far, revising both spatial relation and concerned objects.

One important aspect to analyze the achieved integration is still to be done: the examination of representational structures in the internal layers. This is the key to understanding the actual achievements of the model. Since we have ensured its performance quality, this will be the next step of our research.

Conclusions

The proposed model performed constantly well over all test sets in every imposed task. It comprehends language in an incremental manner and produced correct syntactic strings, even with only the mutually restricted information available. It showed systematic behavior for different demands and conditions concerning missing or ambiguous information. Notably, all specifics of the behavior resulted from inherent properties and were not imposed artificially. The model integrates different information sources, which enables it to produce context sensitive behavior, to reduce irrelevant information and to draw inferences on possible target states. This leads to effects best described in terms of visually induced linguistic performance and language driven shift of visual attentional focus.

Our retina-like representation format clearly limits the possibilities of representing world situations. But it avoids the theoretical shortcomings of using explicitly assigned relations. We assume that it is an inherent and distinctive feature of connectionist models to extract and assign the relations given in the world by themselves.

Further elaboration of the model will first of all encourage an independent generation of prototype situations. The weakest point of the model at present is the explicitly provided prototypicality, i.e. the supervised training of the situations using pre-structured prototypes. To ensure complete independence of the model, we need to equip it with an unsupervised training algorithm that self-organizes prototype extraction using auto-associative learning mechanisms. This way, the model could construct its own prototypical instantiations, corresponding to the achievement of auto-associative lexical models (e.g., McClelland & Rumelhart, 1985). A further improvement of the expressive and predictive power is the extension of situational complexity. The dimensional aspects of time and causality could be introduced by object movement and interaction. This will subsequently extend the vocabulary and the syntactic complexity of the linguistic input, enabling us to construct richer and more language-like descriptions.

Embodiment theories guided the implementation of the connectionist model. On the other hand, explorations of the inherent systematic performance of artificial neural networks can provide a useful subsidiary explanation base to sharpen some opaque conceptualizations in the embodied cognition literature.

References

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577-660.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal recursion semantics: an introduction. *Research on Language & Computation, 3 (4)*, 281-332.

Coventry, K.R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., & Richards, L.V. (2004). Spatial prepositions and vague quantifiers: implementing the functional geometric framework. *Proceedings of the Spatial Cognition Conference*, Frauenwörth.

Dominey, P.F. (2003). Learning grammatical constructions in a miniature language from narrated video events. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Frank, S.L., Haselager, W.F.G., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition, 110*, 358-379.

Frank, S.L., Koppen, M., Vonk, W., & Noordman, L.G.M. (2007). Modeling multiple levels of text representation. In F. Schmalhofer, & C.A. Perfetti (eds.). *Higher level language processes in the brain: inference and comprehension processes.* Mahwah, Erlbaum, 133-157.

Joyce D., Richards L., Cangelosi, A., & Coventry K.R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. *Proceedings of the 5th International Conference on Cognitive Modeling*, Bamberg.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Mayberry, M.R. (2003). *Incremental nonmonotonic parsing through semantic self-organization*. Doctoral Thesis. University of Texas, Austin.

Mayberry, M.R., Crocker, M.W., & Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual worlds. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Nashville.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology*, 114 (2), 159-188.

Quine, W.V. (1968). Ontological Relativity. *The Journal of Philosophy*, 65 (7), 185-212.

Rohde, D.L.T. (1999). *LENS: the light efficient network simulator*. MIT, Carnegie Mellon University Pittsburg.

Rohde, D.L.T. (2002). *A connectionist model of sentence comprehension and production*. Doctoral Thesis. MIT, Carnegie Mellon University Pittsburg.

Rosch, E. (1978). *Cognition and categorization*. Hillsdale, N.J., Erlbaum.

Steedman, M. (1999). Connectionist sentence processing in perspective. *Cognitive Science, 23 (4)*, 615-634.

Zwaan, R.A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In B.H. Ross (ed.). *The Psychology of Learning and Motivation, 44.* New York, Academic Press, 35-62.