Selective Visual Perception Driven by Cues from Speech Processing

Reinhard Moratz, Hans-Jürgen Eikmeyer, Bernd Hildebrandt,
Alois Knoll, Franz Kummert, Gert Rickheit, Gerhard Sagerer

SFB 360, Universität Bielefeld, Postfach 100131, 33501 Bielefeld

1 Introduction

At present, there is a large number of systems dealing with various aspects of understanding either speech or vision. Such systems are, however, relatively seldom capable of a complex processing that begins with sensory data and arrives by way of subsymbolic processing levels at a symbolic representation. (An overview of current systems is to be found in AAAI-94, Integration of Natural Language and Vision Processing 1994.) Yet the establishment of a connection between perception and the conceptual level is in fact a prerequisite for the modelling of cognitive processes as required for example within the framework of a cognitive semantics (e.g. Johnson 1987; Lakoff 1987; Jackendoff 1983).

In the following, we will present an integrative system which, on the basis of semantic networks, facilitates the symbolic interpretation of sensory data on the one hand whilst providing on the other a uniform representation for visual and linguistic knowledge by means of which an interaction between modules is simplified.

The system is being developed at the University of Bielefeld within the framework of the special research project "Situated Artificial Communicators" (SFB 360). By the term "artificial communicators" we mean formal systems which reconstruct the behaviour of natural communicators in relevant aspects. Since most cognitive skills are situation-dependent, the SFB's research has concentrated on a specific basis scenario. The subject of this scenario is a task-orientated discourse. The assembly of a model aeroplane from construction-kit parts serves here as an illustration (see Figure 2).

As the formalism for knowledge representation, ERNEST has been used (Niemann, Sagerer, Schröder & Kummert 1990; Kummert, Niemann, Prechtel & Sagerer 1993). ERNEST is a semantic network formalism which, besides the general characteristics of semantic networks such as the representation of concepts and the relationships between them, also has a variety of additional features which are highly interesting from a cognitive perspective (Johnson-Laird, Herrmann, Chaffin 1984). The interpretation of sensory data is, for example, seldom unambiguous. Thus cognitive models must be able to interpret robustly both graded and uncertain knowledge near the signal. With ERNEST it is possible to trace competing hypotheses in parallel and to include different areas of knowledge at an early stage in the processing.

In the following, ERNEST's most important characteristics will first be outlined. This will be followed by a more detailed description of those features of ERNEST which seem particularly well-suited to the cognitively adequate modelling of processes of image and speech understanding. Finally, an integrative architecture for both image and speech processing is presented. This architecture makes possible a close coupling of speech understanding processes and attention-driven image processing.

2 Knowledge representation in ERNEST

2.1 The ERNEST formalism

ERNEST (from "ERlangen semantic NEtwork sySTem") is a formalism for the representation of knowledge (Niemann, Sagerer, Schröder & Kummert 1990; Kummert, Niemann, Prechtel & Sagerer 1993). The main influence on the development of ERNEST was the theory of semantic networks as described for example by Sowa (1984, 1991). Important elements of a semantic network are concepts, their attributes and the relations between concepts. These are usually represented as nodes, their internal structures and links between nodes. ERNEST's main task is to interpret sensory data symbolically; in this context these data are primarily visual and acoustic signals. In ERNEST there are three types of nodes:

- A *concept* can represent a class of objects, events or abstract conceptions.
- An *instance* is understood as the concrete realisation of a concept in the sensory data; i.e. an instance is the copy of a concept by which the general description is replaced by concrete values.
- In addition, there are also *modified concepts*. A modified concept represents knowledge which is adapted to a concrete situation of analysis.

Features of a concept, such as the size of an object or the grammatical number of a noun phrase, can be represented by means of attributes. In this way, concepts in ERNEST are given an internal structure. Since the attributes of a concept are sometimes dependent on each other, ERNEST also makes it possible to represent relationships between attributes. In ERNEST, there are the following link types:

- Through the link type *part*, two concepts are connected with each other if one concept is understood as a part of the other concept.
- Another well-known link type is the *specialisation*, with a related inheritance mechanism by which a special concept inherits all properties of the general one.
- The link type *concretisation* connects two concepts to each other if a concept is represented on different levels of abstraction. Thus the visual perception of an ellipse may be a hole or a tyre on a higher level of abstraction.

As mentioned at the outset, the goal of an analysis in ERNEST is the symbolic interpretation of sensory data, i.e. the instantiation of concepts. The creation of modified concepts and instances constitutes the knowledge utilization in the semantic network. For the creation of instances, this process is based on the fact that the recognition of a complex object has the detection of all its parts as a prerequisite. For concepts which model terms only defined within a certain context the instantiation process must proceed in the opposite direction. In this case the context must exist before an instance of the context-dependent concept can be created. In the network language, these ideas are expressed by six problem-independent inference rules. Context-independent parts, contexts, and concretes are the prerequisites for the creation of instances and modified concepts in a data-driven strategy. The opposite link directions are used for model driven inferences. Since the results of an initial segmentation are not perfect, the definition of a concept is completed by a judgement function estimating the degree of correspondence of a part of the signal to the term defined by the related concept. On the basis of these estimates and the inference rules an A*-like control algorithm is applied. For a more detailed description of the network control see (Kummert, Sagerer & Niemann 1992; Kummert, Niemann, Prechtel & Sagerer 1993).

2.2 The representation of cognitive processes

The architecture of a computer system integrating speech and vision processing can be structured according to the following dichotomies. An obvious starting-point is the aforementioned division into visual and linguistic components. Furthermore, a distinction between a long-term memory and a working memory can be made. In ERNEST the long-term memory is represented in the form of concepts and their relationships, i.e. as a semantic network. The instantiation of concepts can be interpreted as the activation of mental entities. During the processing, modified concepts are adapted to the given concrete situation. This constitutes the working memory. Thus a division into a visual and a linguistic working memory, as suggested for example by Baddeley (1983), and their gradual integration into a common mental representation, is given.

In addition, the represented knowledge is also structured in functionally differentiated modules. In the speech components these are for example the areas of syntactic or semantic knowledge. These modules are, however, not isolated, but are on the contrary capable of close interaction. Accordingly information from various sources can be utilised during processing (Altmann & Steedman 1988). In ERNEST, these interactions are supported by a homogeneous representation of the knowledge base. The processing direction between the modules can be data-driven, starting from the perception and arriving via various processing levels at a more or less complete mental representation. Depending on the processing level, model-driven processing, based on knowledge, is also possible. In ERNEST, a flexible strategy facilitates the alternation between data-driven and model-driven processes.

Ambiguities or competing hypotheses which arise during processing are thereafter further processed in parallel. As soon as sufficient evidence has been found for an interpretation (e.g. high activation), this is given preference and expanded further. Thus ERNEST's control makes a (quasi) parallel processing of competing interpretations possible.

ERNEST also distinguishes between attributes and procedures determined by the model, and those which are technically motivated. These features of ERNEST are indispensible for an adequate implementation of cognitive models.

In view of the capabilities described above, we regard ERNEST as a formalism for the interpretation of knowledge which is not only particularly well-suited to the realisation of highly efficient, application-orientated systems, but which is equally suitable as a basis for computer simulations of cognitive processes.

3 Integrative architecture

3.1 The speech understanding component

The speech-processing component in ERNEST is based on a speech-understanding dialogue system for railway information (Mast, Kummert, Ehrlich, Fink, Kuhn, Niemann & Sagerer 1994). The system's aim is to automatically understand spontaneous spoken language and to answer the questions put. A speech recognition system delivers word hypotheses at the interface with the linguistic knowledge base.

The structuring of the knowledge base is orientated towards Winograd's (1983) cognitive speech-processing model, which advocates stratified processing. In this model, the essential processing levels are respectively a syntactic, a semantic and a pragmatic level, all based on a uniform representation.

Since the order of syntactic constituents in spoken German language is relatively free, and only constituents are syntactically stable, no attempt was made to model a complete sentence grammar. The semantic level follows Fillmore's (1968) deep case theory, according to which syntactic-semantic roles are associated with verbs. By means of a verb-orientated, model-driven analysis in ERNEST, concrete expectations as to the argument of a given verb could thus be produced. Using ERNEST's capabilities, the processing strategy in the dialogue system alternates between data-driven and model- driven processes. This facilitates the efficient use of relevant information from both the acoustic data and the linguistic knowledge base.

Figure 1 shows the individual levels of the integrative architecture's speech understanding component, illustrated by examples of some concepts. The first letter of a concept's name indicates the level to which it belongs: P_CUBE is a concept on the pragmatic level, S_OBJECT is a concept on the semantic level, and SY_NOUN is a concept on the syntactic level. Since the speech recognition system can now deliver the full forms of words as hypotheses on the hypothesis level H, there is a so-called word level W, which includes the collective concepts for individual word forms. At present, this word level is structured according to morphological and phonetico-phonological criteria; in the near future it will form a new interface with the speech recognition system. This structuring follows psycholinguistic criteria, so that a computer simulation of lexical access during

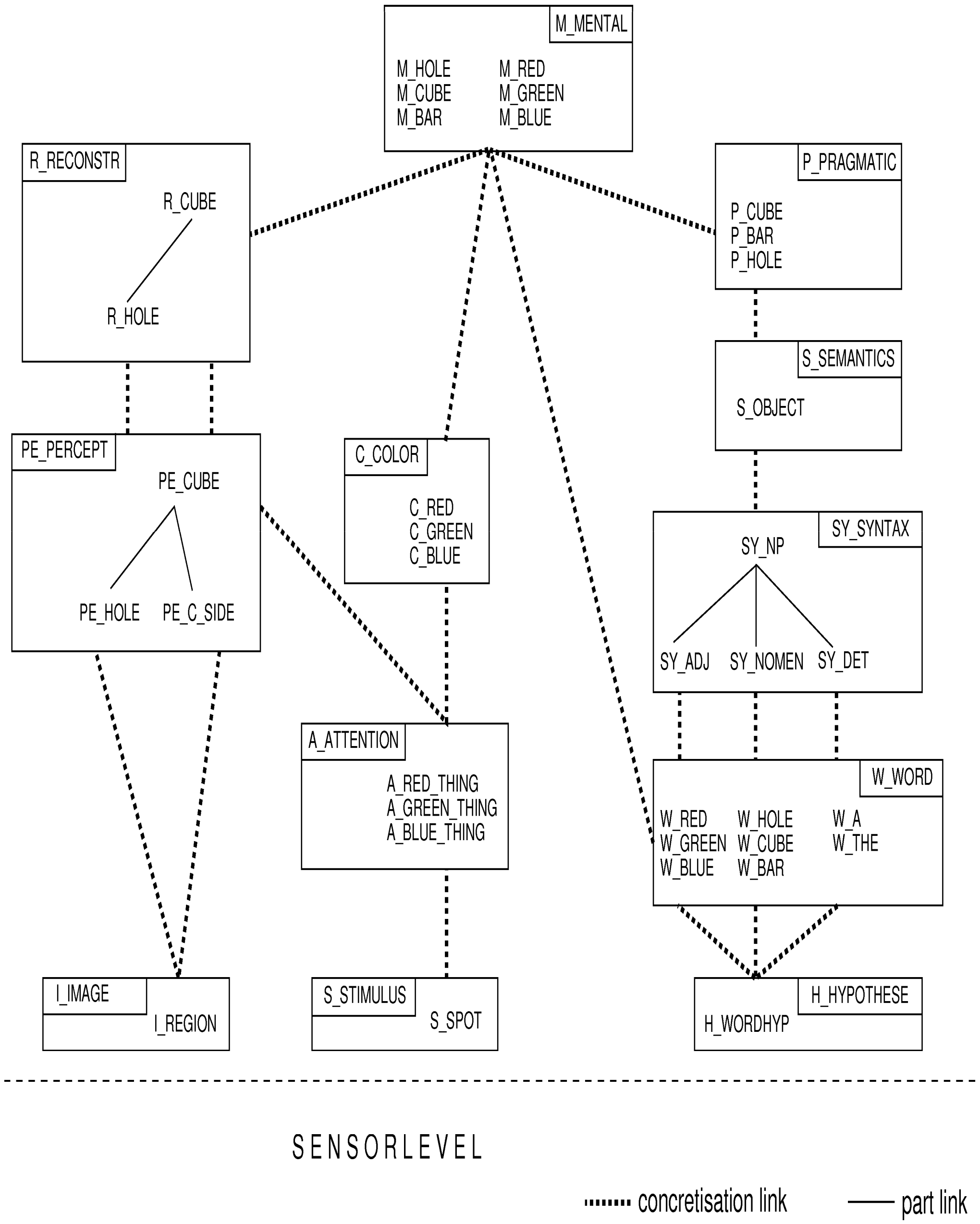

Figure 1 integrative architecture (simplified)

speech production or reception is also possible. For instance, the first syllable of a word is accorded paramount significance, since this is usually what allows the word's semantic content to be indentified early and with a high degree of probability (Marslen-Wilson & Welsh 1978; Levelt 1994; Spivey-Knowlton, Sedivy, Eberhard & Tanenhaus 1994).

3.2 The image understanding component

Model-driven processes are particularly important for cognitively motivated vision processing where an interaction with speech is being aimed at. The latter demands the consideration of special purposes which are derived from the system's situated global requirements (Ballard 1991, Brown 1992).

For this reason, the exclusion of model-driven influences, as often practised in more traditional approaches (e.g. Marr 1982), is no longer appropriate for a comprehensive architecture. In our approach, objects are modelled by means of individual entities which can be robustly detected and which specify the object redundantly. In this, lighting conditions and perspective are taken into consideration on the perceptive level. In Figure 1, this corresponds to level *PE* of the knowledge base. An object corresponds to only a few percepts, since only a small number of topologically differing views can be derived from the contour structure of an object (Koenderink & van Doorn 1979; Rieger 1990). For example, a rhomb-nut is modelled by its upper side, the hole and the two visible sides in front, in a particular spatial arrangement. In ERNEST, a spatial arrangement of this kind can be represented by a relation between attributes within concepts. A three-dimensional reconstruction of the scene is created on subsequent processing levels. In Figure 1, this level is shown by concretisation level *R*.

The interface with the segmentation processes on the signal level (level *I*) is given by contours on the one hand and regions of homogeneous colour and texture on the other. Figure 2 shows segmentation results consisting of ellipses and line segments by intensity edges and also regions by fast colour segmentation.

In selective perception, spatially-oriented attention with low resolution (level *S*) provides an important initial indication for subsequent focussing mechanisms (level *A*). Object-related attention is realised in the *PE* module that has been described above. This architecture allows an interaction of low resolution and colour with subsequent focusing and shape processing. As such, it represents an abstraction and coarsening of biologically motivated architectures which model saccadic eye movements (e.g. von Seelen, Bohrer, Engels, Gillner, Janßen, Neven & Schöner 1994).

3.3 The integration of the components

The integration of the two modalities speech and vision is based on Johnson-Laird's theory of mental models (1980, 1983). "The theory of mental models assumes that they can be constructed on the basis of either verbal or perceptual information" (Johnson-Laird 1980, p.100). Important in this regard is the integrative and coherent representation of objects and facts as well as the cognitive

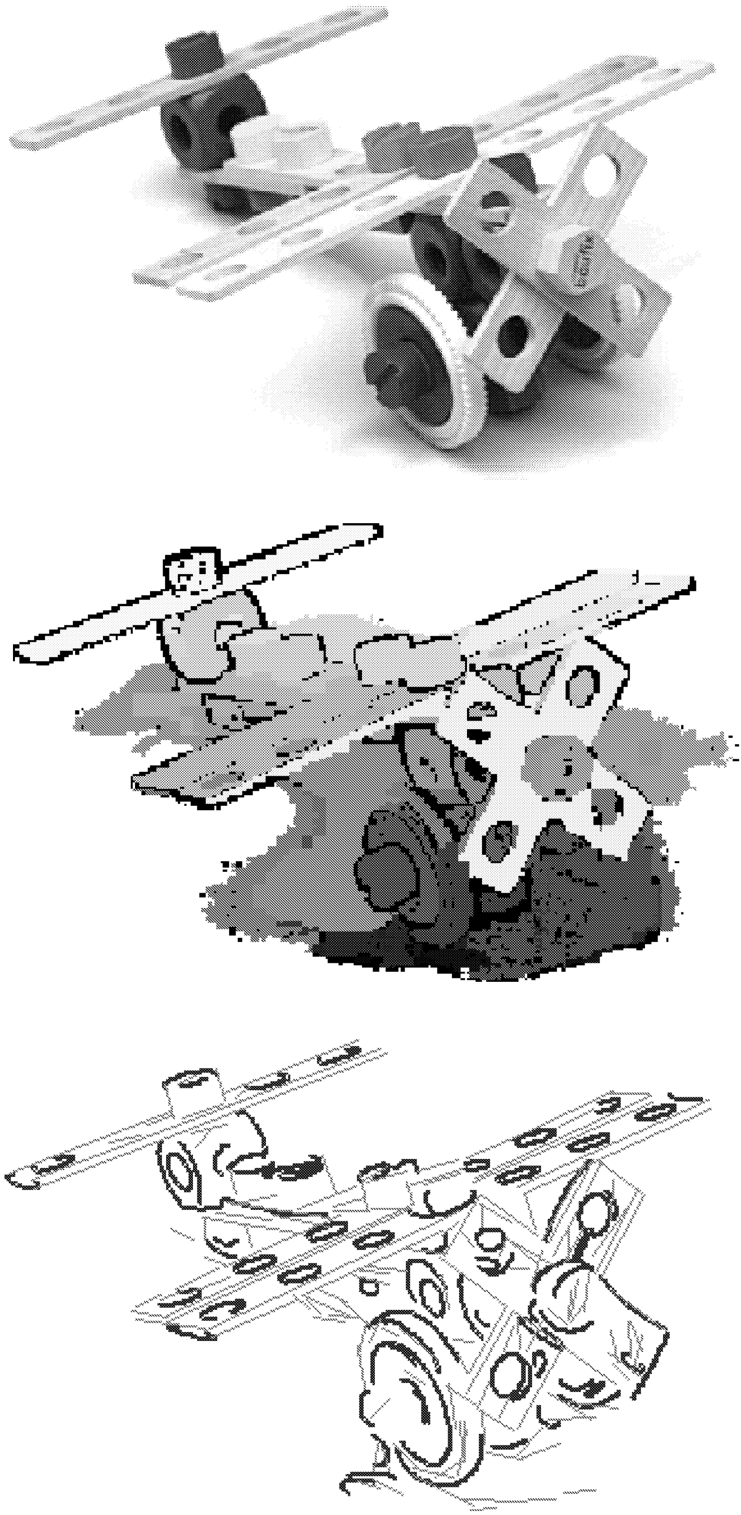

Figure 2 Contour chains and regions

processes based upon such a representation. By means of direct access to various aspects of a concept, it is possible to provide adequate modelling of temporal sequences of cognitive processes. For instance, on the evidence of current psycholinguistic experiments, it seems likely that word recognition may have a direct influence on saccadic eye movements, i.e. on visual processing, even when a word which is heard has not yet been processed through all levels (Spivey-Knowlton et al. 1994).

The integration of the individual components of the knowledge base in ERNEST takes place on a common level of abstraction, which we take as a represention of mental models (level M). In Figure 1, the entire conceptual hierarchy can be seen. A significant characteristic of this hierarchy is that concretisation relations exist not only between adjacent processing levels, but that there are also direct connections between the mental models level and the conceptual level near the signal. By this means, the modelling of an early interaction between the visual and speech components should be possible. The following extract from a dialogue serves to illustrate such an interaction as a sequence of ERNEST inferences:

I: Hast du. So, jetzt nimmst du dir den roten Würfel
K: <pause .> Ja.
I: und die grüne ganz lange Schraube.
K: <pause ..> Mhm.
I: So, den roten Würfel schraubst du jetzt unter den blauen Würfel.

Translated into English:

I: Got it. So, now you take the red cube
K: <pause .> Yes.
I: and the green, very long bolt.
K: <pause ..> Mhm.
I: So, now you screw the red cube under the blue cube.

In addition to the immediate influence which the instructor's directions have on the conscious actions of the constructor, unconscious processes are also set in motion. As mentioned above, experimental data suggest that an utterance on the part of the instructor will influence the eye movements of the constructor. The implication for the modelling is that visual processes already begin during the incremental processing of the verbal instructions. For example, the visual search for red objects in the scene can begin as soon as the word *red* from the first instruction "now you take the red cube" has been understood, before a complete linguistic interpretation of the entire utterance has been completed.

In the integrative system presented here, a concept *W_RED* is instantiated on the word level by the colour adjective *red*, thus giving lexical access. Thanks to direct concretisation links with the syntactic and mental processing levels, modified concepts are laid down simultaneously on the *SY* level as well as the

M level. From the *M* level, a model-driven activation of modified concepts in the visual components now takes place, before the linguistic processing has been completed. A modified concept *C_RED* is activated and following that, a modified concept *A_RED_THING*, which represents red objects in the scene. By laying down a modified concept *S_SPOT*, the red components of the image are emphasised, and red objects in the scene can be directly accessed. If a cube is the only red object in the scene, this can lead to an instantiation of a concept *PE_CUBE* on the perceptive level before the instructor has actually uttered the word cube. The foregoing sequence illustrates by way of example the modelling of interactive vision and speech processing. Due to many degrees of freedom in our model, more detailed empirical data are however still required for a more differentiated modelling.

4 Conclusion

The integrative architecture presented here is based on the requirements arising from the situated integration of speech and vision. This integration requires a homogeneous representation for image and speech understanding. Accordingly, a close interaction of early processes provides the vision module with cues for colour focusing and selective shape processing. We are adapting our present system for a distributed workstation environment based on message passing.

The architecture which has been briefly outlined here offers great potential for further research. Psycholinguistic experiments are needed however, since the process of speech and vision interaction is at present underspecified from the empirical point of view.

Acknowledgement

This work has been supported by the German Research Foundation (DFG) in the SFB 360 project.

Bibiliography

Altmann, G. & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

Baddeley, A. D. (1983). Working memory. *Philosophical Transactions of the Royal Society*, *302*, 311-324.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, *48*, 57-86.

Brown, C. (1992). Issues in Selective Perception. *11th International Conference on Pattern Recognition*, 21-30.

Fillmore, C. J. (1968). A case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (pp. 1-88). London: Rinehart and Winston.

Jackendoff, R. (1983). *Semantics and cognition.* Cambridge, MA: MIT Press.

Johnson, M. (1987). *The body in the mind. The bodily basis of meaning, imagination and reason.* Chicago: University of Chicago Press.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*, 71-115.

Johnson-Laird, P. N. (1983). *Mental models. Towards a cognitive science of language, inference, and consciousness.* Cambridge: CUP.

Johnson-Laird, P. N., Herrmann, D. J. & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin, 96*, 292-315.

Koenderink, J. J. & van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics, 32*, 211-216.

Kummert, F., Niemann, H., Prechtel, R. & Sagerer, G. (1993). Control and explanation in a signal understanding environment. *Signal Processing, 32*, 111-145.

Kummert, F., Sagerer, G. & Niemann, H. (1992). A problem-independent control algorithm for image understanding. *11th International Conference on Pattern Recognition*, 297-301.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Levelt, W. J. M. (1994). On the skill of speaking: How do we access words? *ICSLP 94, International Conference on Spoken Language Processing, September 18-22, 1994, Yokohama, Japan.* Yokohama, Japan: Acoustical Society of Japan, 2253-2258.

Marr, D. (1982). *Vision.* San Francisco, CA: Freeman.

Marslen-Wilson, W. D. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.

Mast, M., Kummert, F., Ehrlich, U., Fink, G.A., Kuhn, T., Niemann, H. & Sagerer, G., (1994). A speech Understanding and Dialog System with a homogenous linguistic knowledge base. *IEEE Transactions on pattern analysis and machine intelligence, 16*, 179-194.

Niemann, H., Sagerer, G., Schröder, S. & Kummert, F. (1990). ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on pattern analysis and machine intelligence, 12*, 883-905.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (ed.), *Semantic information processing.* Cambridge, MA: MIT Press.

Rieger, J. H. (1990). The geometry of view space of opaque objects bounded by smooth surface. *Artificial Intelligence, 44*, 1-40.

Rumelhart, D. E. & McClelland, J. L. (1988). *Parallel distributed processing. Vol.1, Foundations* (8th ed.). Cambridge, MA: MIT Press.

Sowa, J. F. (1984). *Conceptual Structures: Information processing in mind and machine.* Reading, MA: Addison-Wesley.

Sowa, J. F. (Ed.). (1991). *Principles of semantic networks: Explorations in the representation of knowledge.* San Mateo, CA: Kaufmann.

Spivey-Knowlton, M., Sedivy, J., Eberhard, K. & Tanenhaus, M. (1994). Psycholinguistic study of the interaction between language and vision. *AAAI-94 Workshop on Integration of Natural Language and Vision Processing. Twelfth National Conference on Artificial Intelligence.* Seattle, WA: American Association for Artificial Intelligence, 189-192.

Tesniere, L. (1965). *Elements de syntaxe structurale.* Paris: Klincksieck.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Treisman, A. & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 459-478.

von Seelen, W., Bohrer, S., Engels, C., Gillner, W., Janßen, H., Neven, H., Schöner, G., Theimer, W.M. & Völpel, B. (1994). Visual information processing in neural architecture. *Mustererkennung 94. 16.DAGM-Symposium Wien.* Wien: Universität Wien, 36-57.

Winograd, T. (1981). What does it mean to understand language? In D. A. Norman (Ed.), *Perspectives on cognitive science* (pp. 231-263). Hillsdale, NJ: Erlbaum.

Winograd, T. (1983). *Language as a cognitive process. Vol.I: Syntax.* Reading, MA: Addison-Wesley.