Construction of Meanings in Living and Artificial Agents

DIZERTAČNÁ PRÁCA

RNDr. Martin Takáč

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
KATEDRA APLIKOVANEJ INFORMATIKY

Aplikovaná informatika 25-11-9

Vedúci/školiteľ záverečnej práce
Doc. RNDr. Ľubica Beňušková, PhD.

BRATISLAVA 2007

Dedicated to my son Samuel, for being my source of inspiration and joy and also for being a big little disturber of his serious and hard-working father.

Hereby I declare that I wrote this thesis myself with the help of no more than the referenced sources.

Vyhlasujem, že predkladaná práca je mojím originálnym dielom, ktoré som vypracoval samostatne, s použitím zdrojov uvedených v zozname literatúry.

_____

Abstrakt

TAKÁČ, Martin: *Construction of Meanings in Living and Artificial Agents.* [Dizertačná práca]. Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky; Katedra aplikovanej informatiky. Školitel: Doc. RNDr. Ľubica Benušková, PhD. Komisia pre obhajoby: Aplikovaná informatika. Predseda: _____. Stupeň odbornej kvalifikácie: Philosophiae doctor v odbore Aplikovaná informatika. Bratislava, 2007. 171 s.

Predkladaná práca sa zaoberá problematikou povahy a pôvodu porozumenia a významov v jazyku s použitím metodológie výpočtového modelovania. Prináša prehľad existujúcich sémantických teórií, študuje významy u živých organizmov v kontexte evolúcie a empirických poznatkov o akvizícii jazyka u detí. Analyzuje prístup k významom v existujúcich výpočtových systémoch s ohľadom na ich ukotvenosť v interakcii s reálnymi alebo simulovanými prostrediami. Hlavné prínosy práce:

Po prvé, návrh novej originálnej sémantiky založenej na tzv. identifikačných kritériách. Sémantika umožňuje reprezentovať nielen statické objekty a ich vlastnosti, ale aj dynamické zmeny, komplexné situácie a udalosti. Všetky reprezentácie významov (konceptov) možno konštruovať automaticky, extrakciou spoločných čŕt inštancií konceptov v rôznych kontextoch. Teória reprezentácie aj mechanizmy konštrukcie významov sú špecifikované tak rigorózne, že umožňujú počítačovú implementáciu.

Po druhé, návrh dvoch výpočtových modelov konštrukcie interakčne ukotvených významov. V modeli individuálnej konštrukcie významov sa inštancie združujú do kategórií na základe rovnakých interakčných resp. motorických programov (afordancií). V modeli sociálneho učenia významov zameranom na vplyv pomenovávania na tvorbu kategórií sú entity združované do kategórií na základe pomenovania rovnakým názvom. Výsledky experimentovania s oboma výpočtovými modelmi potvrdzujú validitu navrhnutých prístupov k automatizovanej konštrukcii významov.

Po tretie, analýza faktorov ovplyvňujúcich stabilitu významov v medzigeneračnom prenose pomocou výpočtového modelu založeného na iterovanom učení.

Kľúčové slová: Výpočtové modelovanie. Kognitívna sémantika. Ukotvenosť symbolov. Automatizované vytváranie konceptov. Akvizícia jazyka.

Abstract

TAKÁČ, Martin: *Construction of Meanings in Living and Artificial Agents*. [Dissertation thesis]. Comenius University in Bratislava. Faculty of Mathematics, Physics and Informatics; Department of Applied Informatics. Thesis advisor: Doc. RNDr. Ľubica Benušková, PhD. Thesis defense committee: Applied Informatics. Committee chairman: _____. Qualification degree: Philosophiae doctor in Applied Informatics. Bratislava, 2007. 171 p.

This thesis addresses the issues of the nature and origin of understanding and meanings in language with the computational modeling methodology. It reviews formal semantic theories, studies meanings in living organisms within their evolutionary context and takes into account empirical findings about language acquisition by children. It analyzes meanings in existing computational systems with respect to grounding in interaction with real or simulated environments. The main contributions of the thesis follow:

First, we define a new original semantics based on so-called identification criteria. The semantics allows for representation of objects, properties, relations, changes, complex situations and events. All meanings can be constructed by extracting cross-situational similarities among instances of a category. Both the theory and mechanisms of meaning construction are specified rigorously enough to allow for implementation in computational models.

Second, we present two computational models of interaction-grounded meaning construction. In the model of individual category construction, the instances are grouped to categories by common motor programs (affordances), while in the model of social learning, focused on the influence of naming on category formation, entities are considered members of the same category, if they are labeled with the same word by an external teacher. Results of experimenting with both models validate the proposed meaning-formation mechanisms.

Third, we report and analyze simulation results of an experiment focused on the dynamics of meanings in iterated intergenerational transmission.

Keywords: Computational modeling. Cognitive semantics. Symbol grounding. Automated concept formation. Language acquisition.

Foreword

The importance of studying the nature of meanings and mechanisms of their construction is threefold. First, we live in times when human-computer and computer-computer interaction is no longer a science fiction, but a practical engineering problem. We need to design representational formalisms that will allow us to endow machines with ontologies necessary for their successful solving of given tasks and for their mutual co-ordination/communication. The representation must be sufficiently complex to capture peculiarities of physical and social environments, including their dynamical character. In open environments, the ability to learn and autonomously construct useful representation of relevant meanings is crucial.

Second, operationalization of semantic theories and building relevant computational models can help clarify the notion of "understanding" in artificial systems that has been a source of controversy in Artificial Intelligence for a long time, and provide mechanisms for symbol and language grounding.

Last but not least, the computational models can help us better understand ourselves. They can have a backward impact on theories of learning and language development, and on cognitive science in general.

This thesis elaborates the idea that we can only talk about understanding in systems that interact with their environment. It reviews existing computational systems from this point of view, proposes a representation of meanings that can be constructed by interaction and validates the approach by experiments with computational models. Original contribution in comparison to existing systems is analyzed.

I got involved in the problematics of computational modeling of language phenomena coincidentally, after choosing de Boer's and de Jong's IJCAI papers to refer about at an informal seminar at the Institute of Informatics FMFI UK in the autumn 1999. My growing interest in language phenomena has drawn me to cognitive science. My research goals have finally crystallized around the issues of cognitive semantics. All of this has happened in a friendly atmosphere of the Institute and later Department of Applied Informatics. For this, I want to thank all my former and current colleagues.

I am deeply indebted to Ján Šefránek, my friend and teacher. Besides his valuable knowledge, he has always been for me a model of personal and moral integrity and a "good spirit" of our department.

Next, I would like to thank my thesis advisor Ľuba Beňušková for reading lengthy drafts of this text and for her good advices. Despite her being on the other side of the globe, she helped me as she could.

My gratitude goes out to Igor Farkaš for our fruitful discussions and for his friendly spirit, and to Ján Kľuka who saved me whenever I was hopeless at technical problems with my computer.

I am indebted to my students of Cognitive Science, Qualitative Modeling and Simulation, and Communication Training for their questions and comments that deepened my understanding of the field.

Finally, my warmest gratitude goes out to my wife Slávka. Without her loving support and patience, I could not make it.

Contents

Theory of Meaning

Computational Models

List of Figures

List of Tables

Theory of Meaning

Chapter 1

Introduction

This thesis addresses the issue of understanding and meaning, especially how the meaning is constructed/learned in living and artificial systems. In this chapter, we start with motivational and historical remarks. A more formal and structured outline of the thesis will be presented in Section 1.4.

1.1 Understanding as a Benchmark of Intelligence

1.1.1 The Turing Test and the Chinese Room

Can machines be intelligent? This issue has become a subject of heated debates soon after the construction of the first computers. Actually, the name of the established scientific discipline Artificial Intelligence (AI), first used in a grant proposal for 2-month studies at Darthmouth College in 1956,[1] sounds somewhat provocative.

To shift away from nonproductive debates, Alan Turing (1950) proposed an operationalization of the question of the intelligence of machines, the *imitation game*, now known as the *Turing test*. The original Turing version of the test involved three parties (the interrogator and two other persons, one of them trying to help the interrogator, and the other one trying to confuse him),[2] simpler versions have later been proposed in which the interrogator is communicating with a system (human or machine) and has to find out

[1] See e.g. Russell and Norvig (1995, p. 17).

[2] The object of the game for the interrogator was to determine which of the other two was the man and which was the woman solely from communication via a teleprinter or a computer terminal. Now, if a machine could successfully take part of the person trying to confuse the interrogator, it passed the test.

whether the system is a human or a machine.

The ability to communicate in (i.e., understand and produce) natural language indistinguishably from humans has been used as a criterion of (machine) intelligence. The Turing test has elicited a lot of discussions and critique. One of the prominent critics of the Turing test, John Searle, has argued that, solely by observing a behavior (the communication, in this case), one cannot determine whether the system *truly* understands what it is talking about. He proposed a thought experiment, known as the Chinese Room (Searle, 1980):

A person who does not speak Chinese is locked in the room. People outside the room send into the room questions written in Chinese. In the room, there is a box with Chinese characters and a book of rules for manipulating the characters, enabling to produce answers for questions written with Chinese characters. The person in the room composes the answers entirely by comparing the shapes of the characters with those in the box and by using formal rules. Let us suppose that he gets so proficient in manipulating the characters that he gives correct answers to the questions. Nobody outside the room can tell that he doesn't speak a word of Chinese, neither he understands the content of the communication he is participating in. He has produced answers by manipulating uninterpreted formal symbols.

The Chinese Room metaphor has several interpretations.[3] We shall only mention one of them: no computer system or program based on formal manipulation with uninterpreted symbols could truly understand (e.g. Chinese), even if it passed the Turing test. People's minds have semantic contents not reducible to purely syntactic symbols. For true understanding, "something else" is required. One of the ambitious goals of this thesis is to seek for a definition of understanding and provide an answer to what this "something else" could be.

1.1.2 Physical Symbol System Hypothesis

Traditional Artificial Intelligence is strongly connected with the notion of *representation*. Representations are structures that exist within the individual and can be interpreted by the individual itself (Pfeifer and Scheier, 1999). They take the form of symbolic structures that computational processes operate on. A mapping between the external world and internal representations is established via encoding and decoding functions obeying the *law of representation*

[3] And it received at least as much criticism as the Turing test, see for example Cole (2004).

`decode[encode(T)(encode(X1))] = T(X1),`

where `X1` is the original external situation and `T` is the external transformation (Newell, 1990, p. 59).

This characterization has been introduced by Newell and Simon (1976) as the notion of the *Physical Symbol System*. The term "physical" means that symbol systems must obey physical laws and be realized in some physical medium (paper, computer, brain). A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves. A symbolic expression designates an object if, given the expression, the system can either affect the object itself or behave in ways dependent on the object. According to the *Physical Symbol Systems Hypothesis*,

> "a physical symbol system has the necessary and sufficient means for general intelligent action. By 'necessary' we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By 'sufficient' we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity." (Newell and Simon, 1976)

Intelligence, in Newell and Simon's approach, is viewed as symbol manipulation. Physical Symbol Systems Hypothesis characterizes the research program of traditional Artificial Intelligence.

1.1.3 Connectionism

An alternative approach to computation has been inspired by human neurobiology: the brain can be viewed as a massively parallel device composed of millions of richly connected simple processors (neurons). Communication between neurons is carried out by analog signals. The representation of information in the system is distributed and redundant. The overall system is robust against the noise and its performance degradation is gradual in case of damage (Beňušková, 2002a).

The research within this new computational paradigm – *connectionism* has begun with the design of *perceptron* (Rosenblatt, 1958) – a simple formal

computational device modeling a neuron. After overcoming some difficulties with mathematical limits of perceptrons (Minsky and Papert, 1969), the connectionist paradigm has gained an influence in the late 80-ies of the 20th century (Rumelhart et al., 1986). Besides its importance as a computational paradigm (Kvasnička et al., 1997), connectionism has been sucessfully applied in modeling of various cognitive processes and language phenomena (Kvasnička and Pospíchal, 2002; Farkaš, 2005; Rogers and McClelland, 2004).

Both paradigms – symbol and connectionist – are based on the assumption that intelligence is based on specific forms of processing of suitably *represented* information (Kelemen et al., 1992, p. 353). However, they differ in the view on the particular nature of this representation and on the way it is processed.[4]

1.1.4 Nouvelle AI: Intelligence without Representation

A radically different view rejecting any representationalism has been proposed by Rodney Brooks (1991b) in his seminal paper Intelligence without Representation. He criticized the approach of traditional AI as fundamentally wrong. He suggested that we should drop thinking and reasoning, and focus on the interaction with the real world. He proposes an engineering methodology for building artificial creatures, of which he emphasizes two crucial properties (Brooks, 1991a):

Situatedness. The robots are situated in the world – they do not deal with abstract descriptions, but with the here and now of the world directly influencing the behavior of the system.

Embodiment. The robots have bodies and experience the world directly - their actions are part of a dynamic with the world and have immediate feedback on their own sensations.

Brooks proposed the so-called *subsumption architecture* based on a large number of loosely coupled processes that function predominantly in an asynchronous, parallel way. In Brooks's view, the intelligent behavior is an emergent effect of interactions with the environment, i.e. the knowledge is distributed both in the individual's architecture[5] and the environment: "the world is its own best model" (Brooks, 1990).

[4]Much more can be said about differences, advantages and disadvantages of symbolic and connectionist approaches. However, this would exceed the scope of this thesis. An interested reader can see e.g. Fodor and Pylyshyn (1988); Gärdenfors (1997); Farkaš (2005).

[5]A robot is endowed with a new behavior by being added another architectural layer. This approach seems to undermine the role of learning. However, Brooks himself admits

The works of Brooks gave rise to so-called Nouvelle AI based on the *physical grounding hypothesis* (Brooks, 1990):

> "To build a system based on the physical grounding hypothesis it is necessary to connect it to the world via a set of sensors and actuators. Typed input and output are no longer of interest. They are not physically grounded."

After all, this is not in contrast with Searle's conclusion that intelligence is a property of *machines*, i.e. embodied systems causally connected with their environment, rather than disembodied computer programs (Ziemke, 1999).

1.1.5 The Symbol Grounding Problem

The Chinese Room Argument led Steven Harnad to formulate his own version of the problem, known as the *Symbol Grounding Problem*: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?" (Harnad, 1990).

As his own solution to the problem, Harnad suggests a hybrid nonsymbolic/symbolic model of the mind, in which the symbolic functions would emerge as an intrinsically "dedicated" symbol system as a consequence of the bottom-up grounding of categories' names in their sensory (nonsymbolic) representations of two kinds:

1. *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and

2. *categorical representations*, which are learned or innate feature-detectors that pick out the invariant features of object and event categories from their sensory projections.

Elementary symbols are the names of these categories, assigned on the basis of their (nonsymbolic) categorical representations. Higher-order *symbolic representations*, grounded in these elementary symbols, consist of symbol strings describing category membership relations (e.g., "An X is a Y that is Z").

---

that, for more complex tasks, a robot needs to develop internal representations (Brooks, 1991c).

The cognitive tasks involved in symbol grounding are iconization, discrimination, identification and composition. *Iconization* is a process of transformation of analog signals to their iconic projections. *Discrimination* between objects is enabled by the existence of iconic representations, upon which it is performed. *Identification* of an object as a member of some kind or category requires categorical representations.[6]

Connectionism is suggested as a natural candidate for the mechanism that creates categorical representation by extracting the invariant features of sensory projections (icons) paired with feedback about names of the respective categories. Once the taxonomy of elementary categories labeled with names exist, the rest of the symbol strings of a natural language can be generated by symbol *composition* alone (e.g. defining zebra as a "horse with stripes"). This way the hybrid connectionist/symbolic model combines strengths and avoids weaknesses of both approaches: pure symbolic models are weak in connecting symbols to their referents in the world, while connectionist models are week in compositionality (Fodor and Pylyshyn, 1988).

1.2 Understanding in Cognitive Science

1.2.1 Language and Cognitive Science

Although we intend to extend the notion of understanding to also include "understanding the world around", a more frequent notion of understanding is connected to the language.

Language is a phenomenon par excellence in cognitive science (Chierchia, 1999; Rybár et al., 2005). The language competence is a unique property of mankind, distinguishing us from other species. Language processing (parsing and production) involves many non-trivial cognitive processes. The question of evolutionary language origins is still an enigma. The generative paradigm of a linguist Noam Chomsky (1957, 1965) who was inspired by studying the processes of children's language acquisition, has influenced not only cognitive science, but has had a huge impact and applications in computer science, too, namely in the formal theory of languages. Semantics and pragmatics of language is an important field, where cognitive science meets logic and philosophy. Pathologies of language reveal important aspects of the functioning of the brain.

Studying these phenomena reveals a lot about the human mind; that is why linguistics, together with psychology, philosophy, artificial intelligence,

[6]We will talk about the differences between discrimination and categorization in more detail in Section 4.4.5.

neuroscience and anthropology, is considered one of the foundational disciplines of cognitive science (Sloan, 1978; Simon and Kaplan, 1990).

1.2.2 Synthetic Modeling Methodology

The methodology of cognitive science draws on the methodologies of its subdisciplines: theoretical analyses (philosophy), empirical research and experiments (psychology), brain imaging methods and studies of the effects of brain lesions (neuroscience), historical, comparative and field research (linguistics and anthropology), and computational modeling (artificial intelligence). Computational, or synthetic modeling approach can be characterized as "understanding by building" (Pfeifer and Scheier, 1999). It works by creating an artificial system (typically a computer model) that reproduces certain aspects of a natural system (typically results gained by experimental empirical research). The focus of interest shifts from reproducing the results of an experiment to understanding why the results come about,[7] inasmuch as construction of the model involves implementation of hypothesized internal mechanisms that should lead to the particular results. This approach is extremely productive and its advantages include:

- the necessity of detailed, rigorous and operational specification of all terms and mechanisms (misty verbal descriptions are not enough),
- direct verification of predictions,
- a possibility to control internal parameters, which are not controllable in empirical research,
- a possibility to reproduce historically remote processes "in silico" or perform research that would be unethical if performed with human subjects (e.g. linguistic deprivation, damaging neural circuits, etc.).

In this thesis, we will adopt the synthetic modeling approach to study the processes of language acquisition, meaning construction and understanding, and their mutual relations.

1.3 Practical Applications

Besides providing a deeper insight into human cognitive functioning, computational models of language origins, acquisition and processing have im-

[7] Reproducing the empirical results is still an important aspect, because of the validity issue: if the simulated behavior does not fit the observed behavior of the modeled natural system, the model is not valid.

portant practical application in human-computer interaction, as well as for communication of artificial agents operating in heterogeneous and open environments, such as the world wide web. The ability to learn an existing language or to build a new one from scratch, together with construction and mutual negotiation of meanings is crucial in the areas where all possible meanings cannot be anticipated in design-time. Automated negotiation of meanings of terms is an important research issue also in EU terminology unification efforts (Micko, 2006; Popper, 2007).

1.4 Outline of This Thesis

The moral of the story told so far is that problems of meaning and understanding rank among crucial problems of Artificial Intelligence and cognitive science and have important practical consequences.

1.4.1 The Goal of This Thesis

The notion of meaning is in the main focus of this thesis. Our principal goals are listed below:

1. Formulate a notion of meaning that should be applicable not only to linguistic humans, but also to preverbal living organisms and artificial systems.

2. In line with the formulated notion of meaning, propose a formal representation of various types of meaning (including objects, properties, relations, dynamic changes, situations and events) in a rigorous-enough way that would allow for computer implementation.

3. Propose mechanisms of autonomous construction/acquisition of meanings and verify them by experimenting with computational models.

1.4.2 Overview of Existing Theories and Methods

Existing formal theories of meaning are reviewed in Chapter 2. Chapter 3 provides an evolutionary view on the origin of meanings in living systems. In Chapter 4, we review and evaluate existing approaches to representation of meanings in artificial systems within the framework of computational modeling methodology. In Chapter 5, we formulate our notion of meaning.

1.4.3 Our Methodology

Our approach is based on computational modeling methodology of "understanding by building". Methodological commitments that our models shall adhere to are proposed in Chapter 6.

1.4.4 Experimental Plan

Our experimental plan is settled in Section 6.2. It consists of

1. rigorous proposal of semantic representation of various types of meanings,

2. proposal of individual and social mechanisms of autonomous construction of such semantic representations,

3. verification of the hypothesized mechanisms by simulations of implemented computational models.

1.4.5 Experiments

The proposed models and experiments form the content of the second part of this thesis. As we have designed several computational models, first we describe their common features in Chapter 7. In Chapter 8, we propose semantic representation of various types of meanings based on cross-situational similarities. Then we present experiments with the models of individual (Chapter 9) and social (Chapter 10) mechanisms of meaning creation. In Chapter 11, we let the meaning-formation process iterate intergenerationally and we analyze its dynamics.

1.4.6 Evaluation

Partial results are discussed at the end of each of the experimental chapters. General discussion including related works, limits of our approach and directions for the future can be found in Chapter 12. Our contributions and evaluation of fulfillment of the goals of this thesis are summarized in the final chapter.

Chapter 2

Formal Theories of Meaning

2.1 Babel of Terms

When we speak, our words and sentences are *about something*, or they *mean* something. What is this something and what does it mean to mean something are the big questions having been studied by philosophers and linguists for many centuries, starting with Aristotle, Aurelius Augustinus and Thomas Aquinas, through Locke and Hume, to de Saussure, Pierce, Frege, Russel, Wittgenstein, Austin, Tarski, Kripke, Montague, Lakoff, Fodor and many others. Rather than a consensus, two thousand years of discussions have brought in a babel of terms, distinctions and notions.

The precise use of terms is important for the purposes of this thesis, because the answers to the questions of understanding and symbol grounding depend on the precise meaning of the terms such as "symbol", "meaning" and others. Before giving our own definitions of terms used throughout the rest of the thesis, we review several established approaches and their notions of basic terms.

2.1.1 Syntax, Semantics, Pragmatics, Semiotics

Syntax is a subfield of linguistics that studies the construction of complex signs from simpler signs (the rules that determine the way sentences are formed by the combination of lexical items into phrases).

Semantics studies aspects of meaning that are expressed in systems of signs (a language, code, or other form of representation).

Pragmatics studies how language is practically used by individuals and communities and how it is interpreted in particular circumstances and

contexts.

Semiotics is the study of *signs* as complex dyadic or triadic relations. It differs from linguistics in that it generalizes from linguistic signs to signs in any medium or sensory modality. Morris (1938/1971) defined semiotics as grouping the triad syntax, semantics, and pragmatics, where syntax studies the interrelation of the signs without regard to meaning, semantics studies the relation between the signs and the objects to which they apply and pragmatics studies the relation between the sign system and its user.

2.1.2 Sense and Reference

Frege (1892/1952) introduced the distinction between *sense* and *reference*.[1] Sense and reference are two different aspects of the meaning of at least some kinds of terms (mainly proper names).

As Frege discovered, a term's reference (the object it refers to) cannot be treated as identical with its meaning. For example, *Hesperos* (an ancient Greek name for the evening star) and *Phosphorus* (an ancient Greek name for the morning star) both refer to Venus, but the astronomical fact that "*Hesperos* is *Phosphorus*" can still be informative, even if the "meanings" of both *Hesperos* and *Phosphorus* are already known. This problem led Frege to distinguish between the *sense* of a word and its *reference*.

Reference of a term is the object[2] it refers to.

Sense of a term is the way in which it refers to its referent.

We can even understand a meaning of words or phrases that have no referent, such as "the biggest integer", or "Cinderella". Hence, it is safer to define a sense as an individuating description (that can be understand with or without a reference), rather than as the mode of presentation of the reference.

It is interesting to note that Frege considered sense to be objective and distinguished it from a subjective *idea*:

[1] In German original, Sinn und Bedeutung.

[2] Some authors distinguish the reference from a referent, whereas a reference is the relation between words (nouns or pronouns) and objects that are named by them, while the object which is named by a reference, or to which the reference points, is the *referent* of the word.

> "If the reference of a sign is an object perceivable by the senses, my idea of it is an internal image, arising from memories of sense impressions which I have had and acts, both internal and external, which I have performed. ... This constitutes an essential distinction between the idea and the sign's sense, which may be common property of many and therefore not a part of a mode of the individual mind. ... whereas in the case of an idea one must, strictly speaking, add to whom it belongs and at what time. The reference of a proper name is the object itself which we designate by its means; the idea, which we have in that case, is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself. The following analogy will perhaps clarify these relationships. Somebody observes the Moon through a telescope. I compare the Moon itself to the reference; it is the object of the observation, mediated by the real image projected by the object glass in the interior of the telescope, and by the retinal image of the observer. The former I compare to the sense, the latter is like the idea or experience. The optical image in the telescope is indeed one-sided and dependent upon the standpoint of observation; but it is still objective, inasmuch as it can be used by several observers. At any rate it could be arranged for several to use it simultaneously. But each one would have his own retinal image." (Frege, 1892/1952).

2.1.3 Connotation and Denotation

The sense–reference distinction is commonly confused with that between connotation and denotation. The connotation–denotation distinction is commonly associated with the philosopher John Stuart Mill.

This distinction is applied mainly to words expressing properties, i.e. predicates such as red, dog, bachelor, rather than naming individuals, so the difference between the two distinctions can be hard to see.

Connotation of a predicate is the concept it expresses, or more often, the set of properties that determine whether an individual falls under it.

Denotation of a concept is the actual collection of entities that fall under it.[3]

[3]In poetry, the terms denotation and connotation are used with different meaning: Denotation is the literal meaning of a word, and connotation is the suggestive meaning of a word.

For example, the connotation of *bachelor* is "unmarried adult man", while its denotation is all the bachelors in the world.

2.1.4 Intension and Extension

Some contemporary philosophers use the terms intension and extension for connotation and denotation respectively.

Intension of a concept consists of the ideas, properties, or corresponding signs that are implied or suggested by the concept.

Extension of a concept consists of the things to which it applies.

The extension of monadic concepts or expressions (i.e. those that can be satisfied by a single object) is the set of things it applies to, e.g. the extension of the word *dog* is the set of all dogs in the world. By convention, the extension of a whole statement is defined as its logical value (true or false).

The extension of relational or polyadic concepts (those relating objects to objects) is the set of all sequences of objects that satisfy the concept or expression in question, e.g. the extension of the word *before* is the set of all (ordered) pairs of objects such that the first one is before the second one.

2.1.5 Sign

Sign is a central concept of semiotics. We make meanings through our creation and interpretation of signs. Signs take the form of words, images, sounds, acts or objects, but all these things become signs only if we attribute them a meaning. "Nothing is a sign unless it is interpreted as a sign" (Peirce, 1931-58). And, anything can be a sign as long as someone interprets it as 'signifying' something – referring to or standing for something other than itself (Chandler, 2007). Now we describe two dominant notions of sign: those of the linguist Ferdinand de Saussure and the philosopher Charles Sanders Peirce.

Sign as a Dyadic Relation

A sign was defined by de Saussure (1916/1974) as a relation between two parts:

Signifier is the perceivable form (the sound or written word) which the sign takes.

Signified is the mental[4] concept or an idea it represents.

According to de Saussure, the association between a signifier and the signified is completely arbitrary. Only after a signifier is combined in the brain with the signified, it can mean something and form the sign. Hence, meaning is ultimately the same thing as the sign, and meaning means that relationship between signified and signifier. Saussure's conception of meaning was structuralist and derived from relations between signs themselves, not from any reference to properties of material things.

Despite their being a matter of public convention, signs can only mean something to the individual, e.g. what red means to one person may not be what red means to another.

Sign as a Triadic Relation

Peirce (1931-58) defined a sign as a relation between three parts:

> "A sign (in the form of a representamen) is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the interpretant of the first sign. The sign stands for something, its object. It stands for that object, not in all respects, but in reference to a sort of idea, which I have sometimes called the ground of the representamen."

To summarize,

representamen is the form which the sign takes,

interpretant is the sense made of the sign by an interpreter, and

object is that to which the sign refers.

For example, the traffic light sign for 'stop' would consist of: a red light facing traffic at an intersection (the representamen); vehicles halting (the object) and the idea that a red light indicates that vehicles must stop (the interpretant) (Chandler, 2007).

It is important to note that Pierce's notion of a sign is not an absolute or ontological property of a thing, but rather it is a relational, situated and interpretive role that a thing can have only within a particular context of relationships. What constitutes a sign for one observer (interpreter), can be

[4]Note that the signified is a mental construct of a thing rather than the thing itself.

Table 2.1: Different nomenclatures for the triadic sign relation.

Author	Terms		
Peirce (1931-58)	representamen	interpretant	object
Nöth (1990)	sign vehicle	sense	referent
Ogden and Richards (1923)	symbol	thought/reference	referent
Steels and Kaplan (2001b)	form	meaning	referent

just a useless or imperceptible noise for another one, depending on the interpreter's embodiment, society and the history of interactions. A particular interaction between the representamen, the object and the interpretant is referred to by Peirce as (an act of) semiosis.

In relation to de Saussure, the representamen is similar to signifier and the interpretant corresponds to signified. The object is missing in de Saussure's model.

Other sign nomenclatures

The triadic notion of sign is sometimes referred to as "the semiotic triangle". However, different authors use different nomenclature for the vertices of the triangle, see Table 2.1.

Types of Sign

Based on the degree of naturalness of the relation in which the representamen refers to its object through a particular interpretant, Pierce distinguishes three modes/types of sign relations (Chandler, 2007):

Index/indexical – a mode in which the signifier is not arbitrary but is directly connected in some way (physically or causally) to the signified. This link can be observed or inferred: e.g. natural signs (smoke, thunder, footprints, echoes, non-synthetic odours and flavours), medical symptoms (pain, a rash, pulse-rate), measuring instruments (weathercock, thermometer, clock, spirit-level), 'signals' (a knock on a door, a phone ringing), pointers (a pointing index finger, a directional signpost), recordings (a photograph, a film, video or television shot, an audio-recorded voice) and indexical words (*that*, *this*, *here*, *there*).

Icon/iconic – a mode in which the signifier is perceived as resembling or imitating the signified (recognizably looking, sounding, feeling, tasting or smelling like it) - being similar in possessing some of its qualities:

e.g. a portrait, a cartoon, a scale-model, onomatopoeia, metaphors, sound effects in radio drama, or imitative gestures.

Symbol/symbolic – a mode in which the signifier does not resemble the signified but which is fundamentally arbitrary or purely conventional - so that the relationship must be learnt: e.g. language in general (plus specific languages, alphabetical letters, punctuation marks, words, phrases and sentences), numbers, morse code, traffic lights, national flags.

2.1.6 Symbol

For Pierce, a symbol was a sign with completely arbitrary link between a representamen and its object. Saussure avoided referring to linguistic signs as 'symbols' at all. However, most nowadays linguists would agree that language is a symbolic sign system.

A definition of a symbol system, mostly used in computer science and AI was introduced by Harnad (1990), based on Newell and Simon (1976) and others:

> "A symbol system is a set of arbitrary 'physical tokens' scratches on paper, holes on a tape, events in a digital computer, etc. that are manipulated on the basis of 'explicit rules' that are likewise physical tokens and strings of tokens. The rule-governed symbol-token manipulation is based purely on the shape of the symbol tokens (not their 'meaning'), i.e., it is purely syntactic, and consists of 'rulefully combining' and recombining symbol tokens. There are primitive atomic symbol tokens and composite symbol-token strings. The entire system and all its parts – the atomic tokens, the composite tokens, the syntactic manipulations both actual and possible and the rules – are all 'semantically interpretable': The syntax can be systematically assigned a meaning e.g., as standing for objects, as describing states of affairs."

2.1.7 Meaning

Meaning is the content carried by the words or signs exchanged by people when communicating through language. Communication of meaning is the main purpose and function of language. In semiotics, the meaning of a representamen (signifier) is its place within a particular sign relation.

Meaning is the central issue of semantics. However, different semantic theories give different answers to the questions of the nature and origin of

meanings and the relation between language and meanings. In the following sections, we review the most influential semantic theories.

2.2 Functionalist Semantics

The basic question is whether there are some kinds of objects – physical or mental – that are the meanings of linguistic expressions. Semantics answering 'yes' to this question are *referential*; semantics answering 'no' are *non-referential*. First we review a non-referential *functionalist* tradition within the philosophy of language.

2.2.1 Meaning in Use

According to (late) Wittgenstein (1953), meaning of words cannot be defined by reference to the objects or things which they designate in the external world nor by any ideas or mental representations that one might associate with them, but rather by how they are used in communication. The language is primarily about action in the real world; hence, meaning is more in the realm of pragmatics than pure semantics. The meaning of a linguistic utterance simply is its communicative function. Wittgenstein likens the use of language to a *game*: meaning something in language is like making a move in a game according to some rules.

2.2.2 Speech Acts

This idea was further elaborated by Austin (1962). In communication, the speaker can have a variety of goals: drawing the hearer's attention to something, describing something, giving information, asking a question, making a request or giving an order. By speaking, he is actually *doing* something – performing a *speech act* (Searle, 1969).

A speech act has an illocutionary force, e.g. directing someone to do something.[5] In children's utterances, realizations of nine types of primitive speech acts can be distinguished: labeling, repeating, answering, requesting (action), requesting (answer), calling, greeting, protesting, practicing (Dore, 1975).

Interestingly enough, speech acts theory has been a theoretical base for a formal specification of communication semantics[6] in Agent Communication

[5]The illocutionary aspect of a speech act should be distinguished from its perlocutionary effect, which is what it brings about, e.g. the doing of the thing by the person directed.

[6]See also Parunak (1996).

Language (ACL) developed by Foundation for Intelligent Physical Agents (FIPA),[7] the eleventh Standards Committee of the IEEE Computer Society.

2.3 Realist Semantics

In referential semantics, linguistic meanings are some objects. Concerning the nature of these objects, the fundamental distinction should be made between the *realist* and *cognitive* (or *conceptualist*) approaches. In the realist approach, meanings are some entities "out there" in the world. In the cognitive approach, meanings are mental entities "in the head".[8]

2.3.1 Extensional Semantics: Meanings are Objects in the World

Meaning in the extensional type of realist semantics is built upon relations to objects in the "world", or formally in a model structure M. Names are mapped onto particular objects – elements of M, and predicates are mapped onto sets of objects or relations in M.[9] By composition, sentences are mapped onto truth values. The corollary of this approach is that meanings are objective and independent of understanding of particular users.

The foundations of extensional semantics were laid by Frege, further developed in the truth theory of Tarski (1933). Limits of the set-theoretic approach were discovered soon. Some predicates did not fit well the extensional definition of meaning, for example, the meaning of *small* cannot be the set of all small things, because an emu is a bird, but a small emu is not a small bird (Gärdenfors, 2000, p. 61).

2.3.2 Intensional Semantics: Meanings are Mappings to Possible Worlds

To remedy the problems of extensional semantics, Kripke (1959, 1963), Montague (1974) and others developed so-called *intentional semantics*. In intensional semantics, elements of a language L are mapped to a set of *possible*

[7] `http://www.fipa.org`

[8] Cognitive semantics should not be confused with the Language of Thought hypothesis of Fodor (1975) stating that people understand language by translating sentences into propositional structures in the internal language called "Mentalese" amenable to rule-driven inferences. When it comes to the semantics of the Mentalese's syntactic structures, Fodor is a realist in that he relies on references in the external world and truth conditions.

[9] See also discussion on monadic and polyadic concepts/predicates in Section 2.1.4.

worlds instead of a single world. A proposition can be defined as a function from possible worlds to truth values, that determines the set of worlds where the proposition is true.

The intensional semantics has been criticized for:

1. being counterintutive (Bealer, 1989),
2. not supporting inductive inferences (Goodman, 1955; Gärdenfors, 2000),
3. having difficulties with expressing an antiessentialistic doctrine (Stalnaker, 1981),
4. that the model-theoretic definition of properties does not and cannot work as a *semantic* theory of properties (Putnam, 1981).

2.4 Cognitive Semantics: Meanings are Mental Entities

The persuasion that meanings are ideas can be traced back to classical empiricists such as David Hume and John Locke. In modern times, Eleanor Rosch and George Lakoff were the first pioneers of the cognitive approach to semantics.

2.4.1 Prototypes and Basic-Level Categories

The work of Rosch (1978) provided empirical evidence against classical Aristotelian view that categories can be characterized by necessary and sufficient conditions. Rosch discovered that category membership is graded, some members are better examples of the category than others, and some categories even have fuzzy boundaries. The best examples of a category are called *prototypes*.

People categorize at different levels, of which the *basic level* has a special status. It is the most general level, at which a common perceptual image and a common motor program can be created for members of a category, and the level with the highest intra-cluster similarity and inter-cluster distinctiveness. Basic-level categories support inductive inferences, i.e. deriving further properties of objects from their membership in a category. When adults speak to children, they tend to use words for basic-level categories and these words are understood and acquired by children first (Rosch, 1978). The level below the basic one is called *subordinate* and the one above it is called *superordinate*. For example, *animal* is a superordinate category of

the basic-level category *dog*, and *bulldog*, *setter*, *terrier* are its subordinate categories.

2.4.2 Embodied Meanings

George Lakoff (1987) used empirical findings of Rosch and others as an argument against *objectivist* view on cognition and meanings that can be characterized by the following statements: Thinking is independent of the body; it is a mechanical symbol manipulation similar to computer algorithms. Symbols (such as words and mental representations) get their meaning by correspondence with the real world; they are objective and independent of the body, perception and neural system.

Objectivist paradigm is based on the above mentioned Aristotelian view on categories existing in the world independent of an observer. Lakoff is a proponent of the opposite *experientialist* view: Categories are not objectively "out there" in the world. They evolve as learned concepts co-determined by bodily experience, mind and culture. For example our categorization of colors is based not only on wavelengths of the light, but also on properties of our neurobiology of seeing and on cultural conventions (Berlin and Kay, 1969). Meaning is not an objective truth, but a subjective construct. Conceptual categories will not be identical for different cultures, or even for different individuals in the same culture.

We organize our knowledge by means of structures called *idealized cognitive models* (ICM). For example, the meaning of the word *Tuesday* can be defined relative to an ICM that includes the natural cycle of the movement of the sun, the notion of the end of one day and the beginning of the next, and a seven-day calendrical cycle (Lakoff, 1987). Similarly, the concept of *weekend* requires a notion of a cycle of work week of five days followed by a break of two days. This model is idealized, because seven-day weeks do not exist objectively in nature, but they are created by human beings in some cultures (different cultures can have different calendrical systems).

Some ICMs have a propositional structure, i.e. they specify elements, their properties and mutual relations, others have an image-schematic structure. *Image schemas*, e.g. Containment, Source-Path-Goal, Center-Periphery, are embodied prelinguistic structures of experience stemming from recurring patterns in our bodily interactions. ICMs can be combined to radial structures, transformed, or used in metonymic and metaphoric mappings.

In metonymic mapping, part of an ICM is used to stand for the whole. Metaphors refer to the understanding of one conceptual domain in terms of another (by mapping some substructure of an ICM to the corresponding structure in another ICM). They typically employ a more abstract concept

as their target and a more concrete or physical concept as their source, e.g. THINKING IS MOVING, AFFECTION IS WARMTH, TIME IS MOTION, etc. (Lakoff and Johnson, 1980). In relation to the theory of meaning, concrete terms are directly grounded in our bodily experience via image schemas and basic-level concepts, and meaning of abstract terms is build from more concrete meanings via metonymic and metaphoric mappings.

2.4.3 Perceptual Symbol Systems

The idea that human conceptual system is grounded in the bodily experience and neural system can also be found in the work of Barsalou (1999). He emphasizes the role of perception and the brain's modality-specific systems in construction of representations. The representation is modal and has the form of perceptual symbols and simulators:

> "During perceptual experience, association areas in the brain capture bottom-up patterns of activation in sensory-motor areas. Later, in a top-down manner, association areas partially reactivate sensory-motor areas to implement perceptual symbols. The storage and reactivation of perceptual symbols operates at the level of perceptual components – not at the level of holistic perceptual experiences. Through the use of selective attention, schematic representations of perceptual components are extracted from experience and stored in memory (e.g., individual memories of *green*, *purr*, *hot*). As memories of the same component become organized around a common frame, they implement a simulator that produces limitless simulations of the component (e.g., simulations of *purr*)." (Barsalou, 1999)

Analogical simulators develop for aspects of proprioception and introspection. These simulators form a basic conceptual system; abstract concepts are grounded in complex simulations of combined physical and introspective events. Barsalou claims that, while the proposed conceptual system stays inherently modal, it is a fully functional conceptual system that also supports productivity, propositions, and abstract concepts.

2.4.4 Neural Theory of Language

Other researchers focused even more on connection between language understanding and neural activity and processing. Correlations between activations of certain neural structures and perceiving, performing, imagining or

talking about the represented content have been found (Pulvermüller, 1999; Rizolatti et al., 1996; Beňušková, 2005).

In 1990, Jerome Feldman and George Lakoff from UC Berkeley set up a research project L_0 focused on language learning from picture-sentence pairs. Later (1997) they extended the scope of the project to biologically plausible computational modeling of all aspects of language. The project (and the research group) was renamed to the Neural Theory of Language (NTL).[10] Their effort is aimed at explaining how brain functions (including emotion and social cognition) work together to understand and learn language. They study this complex questions on multiple layers (ordered top-down in the list):

Cognition and Language – cognitive mechanisms, linguistic phenomena (spatial relations, metaphor, aspect, epidosic memory, frames, constructions).

Computation – formalisms, data structures, algorithms (executing schemas, feature structures, maps, belief nets).

Structured Connectionism – distributed networks of units (temporal binding, recruitment learning).

Computational Neurobiology – models of neuronal structures and processes.

Biology – biological and neurophysiology structures and processes (fMRI imaging).

The focus of the group is mostly on the required representations and computations, less on neurobiology and the role of particular brain areas.[11]

The NTL assumption is that people understand narratives by subconsciously imaging (or simulating) the situation being described. When asked to grasp, they enact it. When hearing or reading about grasping, they simulate grasping, being grasped, or watching someone grasp.

The main practical results of the group were a computational model of learning the meaning of spatial relations from named pictorial examples (Regier, 1992), a computational model of learning simple action verbs from labeled examples of structured event descriptions – VerbLearn system of Bailey (1997), a model of metaphorical understanding (through embodied

[10] `http://www.icsi.berkeley.edu/NTL`

[11] A reader interested in cognitive neuroscience may want to see the major reference book of the field edited by Gazzaniga (1999). Recent discoveries in the interplay between neural mechanisms and genetics can be found in Beňušková and Kasabov (2007).

simulations) of events in newspaper articles – KARMA system of Narayanan (1997).

The research of embodied meaning of words was extended to larger units such as sentences, modeled by so-called *Embodied Construction Grammar* – ECG (Bergen and Chang, 2003). In ECG, every linguistic unit is a form-meaning pair, where the form part expresses syntactic constraints and the meaning part is a computational-level description of embodied conceptual structures such as image schemas or action simulators (see also Section 4.6.2). The overall NTL research is summarized in the recent book of Feldman (2006). We review several other neurally plausible cognitive models related to the acquisition of language and meaning, proposed outside the NTL group, in Section 4.7.

2.4.5 Conceptual Spaces

Now we turn to cognitive semantics elaborated more on the level of representational structures, without direct connections to neural realization.

Gärdenfors (2000) characterizes cognitive semantics by the following six tenets:

1. Meaning is a *conceptual structure* in a cognitive system (not truth conditions in possible worlds).

2. Conceptual structures are *embodied* (meaning is not independent of perception or of bodily experience).

3. Semantic elements are constructed from *geometrical* or *topological* structures (not symbols that can be composed according to some system of rules).

4. Cognitive models are primarily *image-schematic* (not propositional). Image schemas are transformed by *metaphoric* and *metonymic* operations (which are treated as exceptional features within the traditional views).

5. *Semantics* is primary to syntax and partly determines it (syntax cannot be described independently of semantics).

6. Concepts show *prototype* effects (contrary to the Aristotelian paradigm based on necessary and sufficient conditions).

The first tenet implies that language understanding cannot be managed by any isolated language module, but it is an integral part of the very same

conceptual system that serves for reasoning, orientation and acting in the world.

The fifth tenet is in opposition with the Chomskian tradition within linguistics (Chomsky, 1957, 1965). Within Chomsky's school, grammar is a formal calculus, which can be described via a system of rules formulated independently of the meaning of the linguistic expressions. Semantics is something independent that is added to the grammatical rule system. On the contrary, within cognitive linguistics, semantics is the primary component (which, in the form of perceptual representations, had existed before language was fully developed). The structure of the semantic schemas puts constraints on the possible grammars that can be used to represent those schemas (Gärdenfors, 2000).

The ideas of the second, fourth and sixth tenets have already been discussed within the above-mentioned approaches to cognitive semantics. The main contribution of Gärdenfors is in elaborating the third tenet. As a framework for a geometric structure used in describing a cognitive semantics, he proposes the notion of a *conceptual space*.

A conceptual space consists of a number of *quality dimensions* such as color, pitch, temperature, weight, and the three ordinary spatial dimensions. The quality dimensions are endowed with certain topological or metric structures; some quality dimensions can have a discrete structure. Dimension that have been vital for the survival of humans seem to be *innate* (genetically evolved) and hardwired in our nervous system. Other dimensions are *learned* and some of them are *culturally dependent* (e.g. our linear conception of *time* in contrast to circular conception in some cultures). Finally, some quality dimensions are introduced by *science*, e.g. the distinction between *temperature* and *heat*, or between *weight* and *mass*.

Dimensions of a conceptual space correspond to attributes of represented objects. Because not all attributes are relevant to all represented entities, dimensions are organized in *domains*.

Conceptual spaces are construed in such a way that representations of similar objects are geometrically close to each other. A particular object is represented as a point (vector of coordinates) in a subspace of one or several domains; the similarity between two objects is inversely proportional to the distance of their point representations in the conceptual space (for the distance to be evaluated, objects must share some domains or a subspace with a common metric).

Representation of natural categories is based on the *convexity assumption*: if two points represent objects that are good examples of a category, then any point in between them must also be a good example of that category (Gärdenfors, 2000). Hence, natural concepts are represented by convex

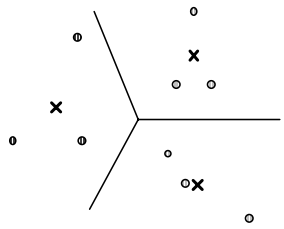

Figure 2.1: Voronoi tessellation of space to categories generated by prototypes. Round points represent examples of categories, 'x' points are prototypes computed as centroids of the examples. The picture is taken from Gärdenfors (2000).

regions in the space. Geometrical centroids of the regions correspond to the best examples – *prototypes* of categories (Rosch, 1978).

Categories can be compactly represented by their prototypes. As categories are considered mutually exclusive, the whole space can be tessellated by assigning each point of the space to the category represented by the nearest prototype (see Figure 2.1). The centroids of the categories can be computed and (continuously recomputed) from incoming examples of the category.

2.4.6 Cognition Without Mental Processes

A cognitive theory of representation of Šefránek (2002) is an effort with a declared goal to posit a non-trivial and falsifiable level of analysis of cognition and understanding without the necessity to resort to the brain and neural processes. Rather, this theory focuses on the *contents* of cognition – *meanings*, which are taken to be *external*[12] with respect to mental processes (Šefránek, 2002, p. 208). This theory aspires to be relevant for real (alive) cognitions in the biological world.

The crucial assumption of this approach is that meanings can be separated from language, i.e. they also exist in animals and preverbal infants. In general, the theory applies to some *organisms* situated in some *environment*. The organisms have *needs* and *goals*, which they try to satisfy by performing *actions* (behavior). The organisms possess *representation* composed of *meanings*.

The theory of meanings is built upon the notion of *recognitional*[13] *crite-*

[12]In this respect, Šefránek's position is at least partially objectivist (Šefránek, 2007).

[13]The Slovak expression "*rozlišovacie* kritérium" can be translated as *discrimination*, *identification*, or *recognitional* criterion. With respect to the sense of the original text and indications in Šefránek (2005, p. 163), here we have chosen the translation *recognitional*.

ria. The recognitional criteria are abstractions of the organism's ability to recognize (or distinguish) certain aspects of its (internal or external) environment. However, an ability to distinguish can be referred to as "meaning", only if it satisfies the following conditions (Šefránek, 2002):

A In a particular situation, the organism can distinguish/recognize also entities that are not directly perceivable.

B The process of recognition is potentially non-deterministic.

C Based on observing and experimenting in its environment, the organism is able to *construct*[14] new meanings (recognitional criteria).

D The organism can *infer* new criteria from the existing ones (by reasoning, without the interaction with the environment).

Elementary recognitional criteria recognize *objects* (individuals), natural *kinds* of objects, natural *properties* of objects, and natural *relations* among objects. More complex criteria, constructed from the elementary ones, recognize *situations*, *rules* (types of situations), *goals* (desired situations), *changes* in environment, *plans* (projected changes), *methods* (successful plans), *events* and *types of events*. The construction of the complex criteria is based on the important notion of *transformations* of criteria.

Šefránek (2002) further suggests the way from protosemantics, protoinference and protocommunication of simple organisms, through 2-language, to the full-fledged language with propositional representation and syntax. The ability to understand the complex language is inherently connected with reasoning, more specifically, with hypothetical[15] reasoning. In his more recent paper, Šefránek (2005) focuses on the possibilities of the representation of verb meanings within his theory.

2.4.7 Other Approaches

In the previous sections, we reviewed several types of cognitive semantics, which were inspiring for our work. Other influential theories of cognitive semantics include e.g. *Force Dynamics* of Talmy (2000), *frame semantics* of Fillmore (1982), *mental spaces* of Fauconnier (1985) and *Cognitive Grammar* of Langacker (1987, 1991b). However, these theories are not in the main scope of our work.

See also the footnote 1 in Chapter 8.

[14]Emphasized by us, also in D.

[15]It is the reasoning with incomplete and dynamic knowledge including the possibility to make mistakes and the necessity of knowledge revisions (Šefránek, 2000).

2.5 Enactive Approach

Cognitivist and representationalist views on cognition have been criticized by proponents of *enactivism*, e.g. Maturana and Varela (1987); Varela et al. (1991). Enactivists emphasize the importance of embodiment, situatedness and action to cognition (Pecher and Zwaan, 2005; Gibbs, 2006). Embodied cognition does not represent some beforehand-given world in some beforehand-given mind. Instead, the individual world (*Umwelt*) and the mind (*Innenwelt*)[16] are being continuously reshaped (enacted) during the interactional history of the individual. There is no linear causal line between perception and action; these two processes form a loop that structurally couples the organism and its environment (Maturana and Varela, 1987).

2.5.1 Subjective Worlds of Experience

The roots of the enactive approach can be traced back to an Estonian biologist Jakob Johann von Uexküll who introduced the notion of *Umwelt*. Von Uexküll was interested in how living beings subjectively perceive theirenvironment. The term Umwelt has later been imported to semiotics by Sebeok (1976).

As opposed to an objectively described world, the Umwelt is the subjectively experienced semiotic world of an organism, including all the aspects of the world meaningful for the organism, e.g. water, food, shelter or potential threats. An organism creates its own Umwelt when it interacts with the world, and at the same time it reshapes the world. The features of the world which the subject perceives (*Merkwelt*) and the features which it acts on the world (*Wirkwelt*) together form a *functional circle* (*Funktionskreis*, von Uexküll, 1934/1957). Subjectively constructed internal representation (modeling system) of the world is the individual's *Innenwelt* (von Uexküll, 1909/1985). In this way, each organism constructs and lives within *its own* lifeworld, which follows from the individuality and uniqueness of its history.[17] The mind and the world are inseparable.

Von Uexküll's work influenced a philosopher Martin Heidegger, who added two other terms: *Mitwelt* and *Eigenwelt*. These terms were further elaborated and used by an existential psychologist Ludwig Binswanger (1942/1993).

Umwelt – subjectively perceived biological and physical world around the individual.

[16] The terms Umwelt and Innenwelt were coined by von Uexküll (1909/1985).

[17] "What a rose is will not be the same to a bee and to a human suitor." (Deely, 2001).

Mitwelt – the individual's social world and his/her awareness, perception and experience of others.

Eigenwelt – the individual himself/herself, including its inner psychological reality and a dialogue with oneself.

2.5.2 Affordances

Another source of inspiration for enactivists was the work of Gibson (1979). According to Gibson, we perceive in order to operate on the environment. Our perception was evolutionarily designed for action. Gibson called the perceivable possibilities for action *affordances*. He claimed that we perceive affordance properties of the environment in a direct and immediate way. Affordances include objects that can be manipulated (such as doors that can be opened), things that can be eaten, surfaces that can be walked on, etc.

Enactive knowledge is the one that comes through action and it is constructed on motor skills, such as manipulating objects, riding a bicycle or playing a sport. Simply, it is the knowledge acquired by doing. Human cognition and consciousness are constituted by the enactive structures – structural couplings between the brain, the body and the physical world with which the body interacts (Varela et al., 1991).[18]

2.5.3 Dynamical Systems Perspective

The enactivist position is most strongly expressed in the Dynamical Systems approach to cognition (Tschacher and Dauwalder, 1999). This approach builds on the mathematical theory of dynamical systems (e.g. Arrowsmith and Place, 1990) and its tools and emphasizes the temporal dimensions of cognition and the ways in which an individual's behavior emerges from interactions of brain, body, and environment (Gibbs, 2006, p. 10). Self-organized patterns of behavior emerge as stable states from the interaction of the subsystems. Dynamical systems perspective cuts across brain-body-world divisions.

Some dynamical models reject representations at all, other incorporate representations, but reconceive them as dynamical entities (e.g., system states, or trajectories shaped by attractor landscapes). Representations tend to be seen as transient, context dependent stabilities in the midst of change, rather than as static, context-free, permanent units (Gelder, 1999).

[18]Note that this approach is not far away from that of Brooks (1991b), see Section 1.1.4.

2.6 Constructivism and Symbol Grounding

The dynamical systems perspective is an instance of *radical constructivism*, which claims that knowledge is the self-organized cognitive process that regulates itself (Ziemke, 2001). In this view, knowledge is a subjective construct with indeterminable relation to an ontological reality. Our position is less radical, but nevertheless constructivist.

Internal representations have no intrinsic meaning *per se* (compare Harnad, 1990), but get it via structural coupling with the environment. This coupling has two components: individual and social. The former one, called *Physical Symbol Grounding* (Vogt, 2002), refers to the ability of each individual to create an intrinsic link[19] between world entities and internal representations, while the latter one, called *Social* (or *External*) *Symbol Grounding*, refers to the collective negotiation for the selection of shared symbols and their meanings (Cangelosi, 2006). A corollary of this is that meanings are individually created subjective constructs, but they are attuned to each other collectively.

Philosophically, physical symbol grounding roughly corresponds to cognitive constructivism based on the work of Piaget (1937/1955) in that individuals actively construct their own meanings through cognitive processes, based upon their past experiences and their interactions in the world. Social symbol grounding is close to social constructivism (Vygotsky, 1978), in the sense that individually created meanings are motivated and constrained by the social context.[20]

2.7 Our Terminology

In this chapter, we reviewed the most influential theories of meaning. Some of them used identical terms with different meanings. Now we try to make clear the way we use some of the terms throughout the rest of this thesis.[21]

[19] E.g. in the form of triadic semiotic sign relations, see Section 2.1.5. We emphasize that all components of the triad are individual, situated and contextual.

[20] Sad stories of *feral children* deprived of linguistic input in the first years of their lives suggest that social and linguistic interactions within a critical period are necessary conditions for the child's successful language development. For example, Genie who had been kept locked in her bedroom, treated badly, deprived from linguistic input and punished for her attempts to speak from the age of 20 month till 13 years by her mentally ill father, have never gained full linguistic competence, despite systematic efforts of her new foster parents and therapists after being rescued from her father. The case of Genie was documented by Curtiss (1977).

[21] Because we need to define the very terms such as "meaning", "sense" or "refers to", we realize that, without resort to meta-language, our definitions will be circular (or struc-

Meaning. In a broader sense, e.g. when talking about "construction of meanings", we refer to construction of the whole semiotic sign triad, e.g. the internal representation (interpretant) put in relation with some lexical expression and referring to some object. For the components of the semiotic triangle, we will use Steelsean nomenclature *form-meaning-referent* (see Table 2.1). I.e., in the narrow sense, meaning refers to internal representation of *concepts/categories*.

Concept/category (equivalent to the narrow sense of meaning) is non-verbal internal representation realized as identification criterion (see Chapter 8). Concepts in the form of criteria are used for identifying (kinds of) objects in the world and they can be connected with lexical expressions (forms) via the semiotic sign relation.

Form/Expression is a synonym for Pierce's representamen, e.g. words.

Reference/Referent is Pierce's object. Unlike Frege (1892/1952), we use this term not in the sense of all objects in the world denoted by some word or concept, but as the set of currently present objects denoted by the word or concept in a particular contextual act of semiosis.[22]

Sense is the particular aspect of the referent inherent in the meaning. For example, when labeling a single big black cup on the table with the expression "big black cup", all three words have the same referent in this situation (the particular cup), but they have different senses. The senses of words are usually disambiguated cross-situationally. The sense of a word is encoded in its meaning, i.e. in the concept linked with the word.

"refers to"/"denotes" : In a general sense (such as in these term definitions), we use these terms for form-meaning mappings, e.g. we can say that a word denotes or refers to some meaning. In connection with a particular act of semiosis of the modeled agents, we will use these terms for form-referent and/or meaning-referent mappings too.

turalist). To keep things simple, we assume the reader's good will and common sense.

[22]For more technical definition of reference, see Section 7.5.1.

Chapter 3

The Origin of Meanings in Living Systems

3.1 Phylogeny, Ontogeny, Glossogeny

Human language is a complex phenomenon that has been co-evolving on different timescales (Takáč, 2003a):

Phylogenetic timescale. This is the scale of biological evolution that has shaped genetically encoded physical and cognitive faculties necessary for language production and understanding.

Ontogenetic timescale. This is the scale of individual language acquisition. Children are exposed to linguistic input in the form of externalized utterances of members of their community – *E-language*. This input shapes the individual's internal representation of language – *I-language*, which controls production and interpretation of sentences of the acquired language (Chomsky, 1986).[1]

Historical/Glossogenetic timescale. The glossogenetic[2] scale is the scale of vertical (from parents/caregivers to children) and horizontal (among peer members of a language community) *cultural transmission* of language (see below). Languages are not transmitted as petrified systems, but they themselves undergo changes and evolve (Deacon, 1997; Kirby

[1] According to Chomsky (1980), the large part of the language competence that codes universal grammatical features (*principles*) is innate, see also Kvasnička and Pospíchal (2005). An innate *language acquisition device* (LAD) controls the ontogenetic process of setting a finite number of language-specific *parameters*, based on the linguistic input.

[2] The term "glossogenetic" that refers to the historical timescale over which languages change was coined by Kirby and Hurford (1997).

and Hurford, 2001), e.g. neologisms appear and archaisms disappear, some elements of syntax simplify and get regularized, etc.

Cultural transmission of language is a non-genetic evolutionary process characterized by reproduction, variation and selection in the following sense (Takáč, 2005a):

Reproduction. The evolving structures are preserved in memories of individuals rather than in genes. The transmission of the structures is realized by behavioral imitation rather than inheritance.

Variation. The imitation/acquisition process is noisy and prone to errors (e.g. overgeneralization or overspecialization) and deliberate innovations of speakers.

Selection. The evolving product is the result of (often conflicting) selection pressures of maximization of communication success, minimization of cognitive processing and memory load, temporal effectivity and constraints of sensory-motor apparatus.

The processes on different timescales do not work in isolation, but are coupled and determine each other. The emerging structure of a glosogenetically evolving language is constrained by the ontogenetic process of language acquisition, which is in turn determined by innate learning mechanisms (Briscoe, 2001).

In this thesis, our primary focus is on meanings. Although the language acquisition process can be viewed as a problem of acquiring correct mappings between elements of overt form, such as words, sentences, gestures, etc. and covert meanings (Langacker, 1991a), we rather view language as a system of triadic semiotic signs (see Section 2.1.5). Within this view, we can study links between any two vertices of the semiotic triangle and mutually interacting processes of their establishment.

3.2 Preverbal Meanings

Meanings, as mental concepts, are taken to be inborn (Fodor, 1981; Rybár, 2005), acquired in the course of interaction with the world (Bloom, 2000), or formed by the influence of language itself (Whorf, 1956). If we took an extreme view that all concepts exist in advance, language acquisition would be just a labeling problem – learning the names for existing concepts. If we took the opposite extreme view, no thinking could exist without a language.

We agree with Šefránek (2002)[3] in that at least some meanings, in the form of embodied knowledge coming from perceiving and acting in the environment, are separated from language, because they can be found in prelinguistic organisms both in phylogeny – in animals (Kováč, 2000) and ontogeny – in preverbal infants (Piaget and Inhelder, 1966; Spelke, 1990). We take a stance that some concepts are innate, others emerge in the process of sensory-motor interaction with the environment, and yet another ones are formed or reformed by the influence of the language. In the computational modeling part of this thesis, we account for construction of preverbal meanings by sensory-motor interaction (Chapter 9) and construction of lexical meanings by verbal instruction (Chapter 10). Computational models of the origin of innate meanings are based on evolutionary algorithms (Kvasnička et al., 2000). An example of such a model can be found in Section 4.3.2.

3.2.1 Phylogenetic Precursors: Signifaction

In this section, we focus on the origin of preverbal meanings, i.e. on the origin of the referent-meaning (object-interpretant) relation of the semiotic triangle. The foregoing study concerns individuals situated in an environment, achieving their goals by sensing and acting. In line with Kelemen (1994), we will call such individuals *agents*.[4]

Some scientists, e.g. de Chardin (1956); Goodwin (1978); Kováč (1986), trace/postulate elementary forms of cognition on very deep levels of the phylogenetic tree, even on bacterial, cellular and molecular levels (Kováč, 2006).

Some of the systems that appeared in the course of evolution have persisted, because their structure reflected relevant characteristics of their environment. Biological evolution consists in generation of hypotheses about the nature of the environment, in falsifying these hypotheses, and in maintaining the hypotheses that have not been falsified (Kováč, 1986).

Hence, evolutionary adaptation of organisms can be viewed as a form of phylogenetic learning with knowledge being encoded in their structure. The chances of persisting are higher for organisms that actively explore their environment and adapt to it or adjust it by their behavior. Each organism recognizes[5] the environmental aspects that are significant for itself.

In the most elementary sense, agents attribute meaning to parts of their

[3] See Section 2.4.6.

[4] Although, in this section, we apply the term agents to living organisms, later we will extrapolate the same principles (and apply the same terms) to artificially created systems.

[5] Once the environmental knowledge is built in the structure of the organism, it is a deterministic process of *re*cognition, rather than non-deterministic cognition (Kováč, 1986).

environment by recognizing, via their sensors and actuators, information useful for achieving their goals (Nehaniv, 2000; Šefránek et al., 2007).[6] Protocreation of meanings in the form of recognition of relevant aspects of the environment is called *signifaction* by Kováč (2003).

The simplest cognitive systems only consist of the mechanisms of sensing and actuating their environment. More complex cognitive mechanisms – perception, affection and cogitation – are gradually interjected between sensing and actuating in the process of intercalary evolution (Kováč, 2003).

Thinking appears on the highest stages of evolution as an abstract action – testing of various motor acts without actually involving the muscles.[7] "What-if" thinking – the ability to mentally simulate various scenarios and evaluate their consequences without the necessity to realize them physically increased the survival chances of organisms and provided them with a significant evolutionary advantage.

3.2.2 Cued and Detached Representations

Gärdenfors (1996a) distinguishes between two kinds of representations: *cued* and *detached*. A cued representation must always be triggered by something present in the current situation. An organism reacting to certain states of its environment in certain ways (e.g. eating objects recognized as food and avoiding objects recognized as predators) performs categorization and possesses cued representations of the respective categories. However, these representations are only activated in the presence of their referents. Cued representations observable as non-volitional behavioral reactions are innate and have evolved phylogenetically.

A significant mechanism that enhances the limited memory of an organism consists in putting externalized marks in the environment, for example effluvial marks that help animals in orientation (or a notoriously known knot in the handkerchief as a "don't forget" sign). In these cases, a mark put in the environment later triggers the respective cued representation.

A detached representation may stand for objects and events neither present nor triggered by anything in the current situation of the organism. For example, a chimp looking for a (non-present) twig to reach for a banana possesses a detached representation of a twig and its use.

It is speculated that the appearance of detached representations in phylogeny co-occurs with the development of neocortex (Gärdenfors, 1996b); in

[6] But see the discussion about when recognitional abilities can be referred to as "meanings", Section 2.4.6.

[7] See also Wiedermann (2007) for a computational account on the emergence of higher cognitive functions in an artificial agent.

ontogeny it corresponds to *object permanence* (Piaget and Inhelder, 1966).

Possession of detached representation is a necessary condition for higher cognitive functions such as planning, deception, self-awareness and linguistic communication (Gärdenfors, 1996a). Planning presupposes the organism's ability to "mentally" (i.e. on the level of detached representations) evaluate expected consequences of various behavioral scenarios and choose the sequence of actions that suits best its current goals. Good planning should also take into account consequences of actions of other agents. Deception presupposes representation of other agents as having their own representations, plans, etc. (i.e. sort of a "theory of mind"). Also, a lier must have a representation of how he will be viewed by the deceived agent. In this way, he has a representation of a representation of himself, which is a necessary precursor of self-awareness (Gärdenfors, 1996a; Beňušková, 2002a).

3.3 Linguistic Meanings

The linguistic competence is considered one of the highest cognitive functions. Language is a symbolic sign system that enables externalization and communication of detached representations. Thanks to its detached nature, it enables talking about things not present here and now, even about things that cannot exist physically. According to Gärdenfors (2004), language evolved in order to make cooperation about future goals possible.

However, this seems to create a paradox: cooperation requires socially shared meanings, but each communication participant has its own individual meanings. How can the participants understand each other? There are several answers to this paradox.

First, although individual meanings are not identical, they must be sufficiently similar thanks to similar learning mechanisms and experiences in a shared environment (Steels et al., 2002). If the meanings are not sufficiently similar, the communication ends up in misunderstanding.

Second, the intended meaning of the speaker is inherently ambiguous: it cannot be *transferred*, but it must be *inferred* by the hearer from the pragmatic context. Inference of the meaning is problematic, as it has been stated by Quine (1960) in the famous *Gavagai problem*:

> Suppose we have a linguist observing a native speaker of a foreign language pointing to a rabbit and saying 'Gavagai'. The linguist cannot be sure what 'Gavagai' means, because it could mean 'rabbit', 'animal', 'white', 'fur' and many other things.

Ostensive definition[8] has also been criticized by Wittgenstein (1953). If someone points to two nuts while saying "This is called two", the listener cannot determine whether the word "two" means the number of items, the type of nut, or their color, unless he has already had an understanding of the process and context involved. Another solution is that an ostensive definition can be variously interpreted in every case. In Section 3.3.3, we review several developmental strategies that help children tackle this problem.

Third, common social meanings can be viewed as constantly renegotiated moving equilibria emerging from the process of mutual coordination of individual meanings of language users (Gärdenfors, 2000). This self-organizing process was modeled by Steels (2000), showing how a globally coherent language can emerge from scratch as a result of local interactions of language users. The community of language users was modeled by a multi-agent system, in which agents (simulated or embodied in real robots) played various types of *language games* by picking a topic from the environment and describing it with a chosen linguistic form (see Section 4.4.1). In the course of time, each agent adapted its linguistic behavior according to the history of previous interactions. A positive feedback between the selection of a language form and its success in use resulted in self-organization and the emergence of a coherent lexicon.

3.3.1 Linguistic Determinism

Linguistic meanings are not isolated, but they are interwoven in a conceptual system. This motivated some thinkers, stemming back to Wilhelm von Humboldt (1820/1997) to postulate that language shapes thought. This idea was further developed by Edward Sapir and his student Benjamin Lee Whorf and has become known as *Sapir-Whorf hypothesis* (Sapir, 1949; Whorf, 1956) consisting of statements of linguistic determinism and relativity. Several versions of these statements have been proposed and debated. Basically, the principle of *linguistic relativity* states that different languages mediate different world views and subtle differences in meanings in one language cannot be easily expressed in another language. Strong version of *linguistic determinism* states that language (completely) determines our thought. This version has been a subject of many controversies and is commonly thought to be incorrect now. Weaker version of linguistic determinism, stating that language *affects* our thought, has been supported by many experiments focusing on whether (cultural) differences in extra-linguistic processes correlate with and depend on

[8] An ostensive definition conveys the meaning of a term by pointing out examples. It is usually accompanied with a gesture pointing out the object serving as an example, e.g. defining "red" by pointing out red objects – apples, stop signs, roses.

differences in linguistic structures (Kay and Kempton, 1984). Results of these experiments have confirmed that there are clearly some measurable effects of the language structure on cognition. In experimental cognitive tasks, speakers of different languages have shown different forms of class generalizations, conceptions of time, spatial orientation systems, color similarity judgments, reaction times and priming effects. References and details of experiments can be found in Feldman (2006, Chap. 15).

Observation of a linguistic behavior of an individual can tell us a lot about his/her conceptual system. This methodology can be extended to a common conceptual system of a community. Language use reveals as a specific interpretation of the world shared by a community – *its language world view* (Bartmiński and Tokarski, 1986). Meaning of a linguistic expression and the corresponding conceptual structures behind it can be deduced from the contexts of its use by a linguistic community. This methodology was used e.g. by Lakoff (1987) in his famous case study of *anger* and by Vaňková et al. (2005) for analysis of Czech language world view.

3.3.2 The Influence of Naming on Concept Formation

An aspect of linguistic determinism most interesting for the purposes of this thesis is the influence of language on concept formation process.

In the experiment carried out by Waxman and Braun (2005), the authors show that naming highlights commonalities among objects for infants and help them to organize objects into categories. The experiment was conducted with 12-month-old infants. In the familiarization phase, the infants were offered objects from a given set (e.g. four different animals) one at a time, in random order. Each infant was assigned to one of three conditions (see Figure 3.1). In the *No Word* (control) condition, the experimenter drew the infant's attention to each object but offered no label, saying, e.g. "Look! Look here!" In the *Consistent Noun* condition, she said, "Look! It's a(n) X!", using the same nonce noun throughout the familiarization trial for a given set. In the *Variable Noun* condition, she said, "It's a(n) X!", presenting a different nonce noun for each named object within a given set.

In the test phase, infants in all conditions were simultaneously presented two test objects: a new member of the supposed now-familiar category (e.g. another animal) and an object from a novel category (e.g. a tool). The experimenter held the objects in front of the infant, saying "Look! See what I have?" and then the infant was allowed to manipulate the objects freely for certain time. The total manipulation time for each of the objects was measured. If the infant formed a category, he/she should preferably attend to the test object not in the category and the manipulation time for this

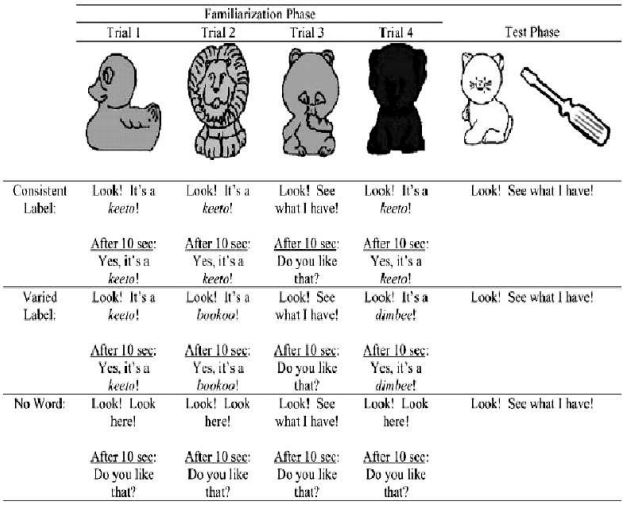

Figure 3.1: An experiment focused on influence of naming on object category formation. A representative sample of the objects and instructions in each condition. The picture is taken from Waxman and Braun (2005).

test object should be significantly longer than for the other. The significant novelty-preference occurred in the *Consistent Noun* condition, while in the *No Word* and *Variable Noun* conditions it did not exceed the chance level. This suggests that, in the two latter conditions, the infants failed to form a category. The experiment focused on superordinate level categories (e.g. animal) because the effect of naming has been most apparent here (at the basic level, e.g. horse, the infants formed categories successfully even in the *No Word* control condition).

In several variants of this experiment, Waxman (2004) tries to clarify *how* exactly naming influences the children's conceptual organization and supports discovery of novel concepts. She compares the effect of naming with a non-linguistic attention-drawing sound and also uses different grammatical forms (nouns and adjectives, e.g. "this is a *blicket*", "this one is *blickish*") to show that even 14-month-old infants start to be sensitive to grammatical clues. Different kinds of words direct the infant's attention to different aspects of the same scene: nouns highlight category-based commonalities, while adjectives highlight property-based ones.

Waxman concludes that naming definitely has the effect on category formation: using distinct names for distinct objects motivates looking for *differences* and supports individuation, while using the same name for distinct objects motivates looking for *similarities* and supports categorization. Booth and Waxman (2002) speculate that object names are salient for infants be-

cause they support communication, predictions about objects in the vicinity and induction of their non-obvious properties, as well as provide better means for reaching desired goals.

A general conclusion that can be drawn from these experiments is that, in children, the processes of language acquisition and meaning formation are not independent, but coupled. In Chapter 10, we present our own computational model of category formation supported by naming. By means of synthetic modeling, we try to verify the hypothesis about the influence of naming on category formation process suggested by the empirical experiments we have just described. The important novel contribution of our model is that meanings of words are not only whole objects, but also their properties, mutual relations and dynamic changes in time.

3.3.3 The Inference of Meanings

In bootstrapping a language from scratch, it is particularly important to establish shared meanings of referential expressions, such as names, nouns and adjectives (Gärdenfors, 2004). This happens by referential (labeling) acts that draw attention to objects present on the scene of communication, with the help of non-verbal means such as pointing, gaze following or joint attention (Tomasello and Farrar, 1986). The whole matter is complicated by the fact that a word uttered along with a non-verbal reference to an object can label the object, any of its parts or properties, its superordinate class and many other things (Quine, 1960). Children use several strategies to overcome this problem: they assume that novel words refer to *whole objects* (Markman, 1992), that a novel word cannot name an object that already has a name (the *mutual exclusivity constraint*, Markman, 1992, 1989), that any difference in form marks a difference in the meaning (the *principle of contrast*, Clark, 1987). They also disambiguate meanings by occurrences of their referents in multiple situations (Akhtar and Montague, 1999; Waxman and Braun, 2005).

Chapter 4

Meanings in Artificial Systems

In previous chapter, we provided an evolutionary view on the origin of meanings in living systems. The nature and origin of meanings is also an important issue in artificial systems and various computational models. We have seen that attribution of "understanding" to artificial systems is a very controversial matter and is connected with certain problems (see Section 1.1).

In this chapter, we review several computational models of various aspects of communication, language origins and language acquisition. We compare the representations of meanings in these models with respect to their expressive power, the ability to cope with the Symbol Grounding Problem (Harnad, 1990) and several other related criteria.

For evaluation of models, we will consider the following aspects:[1]

1. Is there any environment in the model that the agents interact with? Is the environment simulated or real (i.e., are the agents embodied in robots)?

2. Are meanings individual, or identical for all the agents? Are they fixed and innate (given beforehand), or constructed and continuously incrementally updated?

3. What is the type of representation of meanings (predicate logics, vectors, neural networks, prototypes, discrimination trees, etc.)?

4. What kinds of concepts can be represented (static objects, properties, changes, actions, events, situations, etc.)?

[1]Not all the listed questions are relevant for each model.

5. Do concepts (categories) have sharp, or fuzzy boundaries? Can they overlap?

6. Does the representation support synonyms, homonyms and hierarchical relations among concepts?

7. Is the representation sensitive to multidimensional concepts based on inter-correlations of attributes?

8. If the meanings are not fixed, what is the driving force of their adaptation (evolutionary fitness, pragmatic feedback, success in discrimination task, observation of coincidences, etc.)?

9. In communication, do agents have a direct access to internal representations of meanings of other agents ("telepathy")?

All the presented models are simplified in some aspects, often deliberately. We need to emphasize that this is not necessarily a fault, as long as the models abstract away from aspects not directly relevant to the research goal of the modelers. Keeping the number of intervening parameters small is important for correct interpretation of simulation results and for identifying causal dependencies in the model.

Our selection of models is not meant to be exhaustive. We have chosen several models to illustrate various possibilities of meaning representation and problems that the chosen representations bring with. In describing the models, we deliberately do not go into more detail than necessary for our purpose. For details, we refer the reader to the original literature.

The reviewed models could be ordered in several ways; we have chosen the order by a type of meaning representation, then by a modeling goal. Anyway, several models span over more than one category.

4.1 Procedural Representation and Rules

4.1.1 Natural Language Understanding

ELIZA

ELIZA was a program written by Joseph Weizenbaum (1966) in 1964-66 at the M.I.T.[2] It carried on conversations with a user in the manner of non-directive Rogerian psychotherapy (Rogers, 1951). As such, it was not supposed to initiate new themes in the conversation and could do with mostly reformulating and mirroring sentences of the user.

[2]Massachusetts Institute of Technology.

Computationally, ELIZA had to parse the input from the user to find the most important keywords to which it should respond, then use rules and templates to transform the input (e.g. turn "I" into "you") into a response that echoed the input statement or asked for more discussion around the keyword. In the case it was unable to identify any keywords in the user's input, it generated a neutral phrase that could sound reasonable.

Evaluation. Although ELIZA got a good public reception initially, and some users even became emotionally attached to it (Weizenbaum, 1976), it is hardly possible to talk about any "meanings" or "understanding" in ELIZA. It had no model of the outside world, nor it used any semantic representation of the conversation. Even there was no goal or overall script of the conversation: ELIZA was purely reactive according to the list of preprogrammed transformational and decomposition rules and it had no learning abilities.

SHRDLU

SHRDLU is a program for understanding natural language, written by Terry Winograd at the M.I.T. AI Laboratory in 1968-70 and described in his dissertation (Winograd, 1971). The program was considered an extremely successful early demonstration of the power of AI.

The program allows a user to converse about a simulated (visualized) simple "block world" consisting of 3D shapes of various colors and sizes, in which SHRDLU acts as a robot with an eye and a magnetic hand. The user can give SHRDLU commands to manipulate the objects in the block world, e.g. "*Find a block that is taller than the one you are holding and put it into the box.*" or ask it questions about the world or about history of its actions and its own "mind", e.g. "*How many blocks are not in the box?*" or "*When did you picked up the pyramid and why?*". The effects of SHRDLU's actions are visualized on the screen. SHRDLU is sensitive to context references, able to resolve ambiguities and make inferences about the world. It is also able to learn simple facts about what kinds of objects the user likes/dislikes, learn definitions of novel words and use them in further interactions, e.g. "A steeple is a pyramid on top of a block." and then "How many steeples are there on the table?".

SHRDLU, written in LISP, consists of mutually interacting modules of syntax, semantics and inference. The inference mechanism, used both for directing the parsing process and for deducing facts about the block world, is realized by the deductive system MICRO-PLANNER (Sussman and Winograd, 1970), which is the core of SHRDLU functioning. Unlike other theorem provers that handle assertions in predicate calculus, MICRO-PLANNER is a goal-oriented procedural language that also interprets procedural knowledge

used for controlling the inference process. SHRDLU maintains an internal model of the world in the form of a collection of MICRO-PLANNER theorems representing the properties and state of the different blocks, e.g. `(#COLOR :BOX #WHITE)`, or `(#AT :B5 (400 600 200))`, or `(#GRASPING :SHRDLU :B2)`.

SHRDLU can perform actions in the block world. Elementary actions `#GRASP`, `#UNGRASP`, `#MOVETO` are directly connected with routines for visualization of the simulated block world. More complex actions are constructed as plans (MICRO-PLANNER programs).

Semantic knowledge about the meanings of words is represented in the form of MICRO-PLANNER programs, e.g. the meaning of "red cube" is the program for verification that the object bound to the variable `X1` is a red block with equal dimensions:

```
(THPROG (X1)
   (THGOAL (#IS $?X1 #BLOCK))
   (#EQDIM $?X1)
   (THGOAL (#COLOR $?X1 #RED)))
```

The meaning of the word "one" is either the number 1, or the program that looks in the history of the dialog for a context reference (depending on the result of syntactic analysis, whether "one" is a noun or a number).

The semantic module is a collection of LISP programs that look at the syntactic structures as well as meanings of words and combine them into MICRO-PLANNER programs. The syntactic parsing of sentences is maintained by the module called PROGRAMMAR, which implements the so-called *systemic grammar* approach to parsing that emphasizes mutual relations of symbolic units and their function in the process of understanding (Rubin, 1973).

Evaluation. We can say that SHRDLU's knowledge is partially represented as predicate-logic clauses and partially as procedures operating on the clauses. The atoms of the representation, such as `:SHRDLU`, `#RED`, `#COLOR`, are symbolic and given beforehand. Limited learning ability of SHRDLU consists in adding new assertions to the world model, based on the user's input and the history of actions in the block world. In the sense that SHRDLU can perform in the simulated world actions specified by linguistic commands and answer correctly questions about the status of the blocks world, it shows what could be called limited understanding of its virtual environment and a small subdomain of the language.

```
INROOM(ROBOT,R1)
CONNECTS(D1,R1,R2)
CONNECTS(D2,R2,R3)
BOX(BOX1)
INROOM(BOX1,R2)

...
```

$$(\forall x \forall y \forall z)[\texttt{CONNECTS}(x, y, z) \Rightarrow \texttt{CONNECTS}(x, z, y)]$$

Figure 4.1: An example of the axiomatic model of Shakey's world Nilsson (taken from 1984).

4.2 Predicate Logic

4.2.1 A Mobile Robot Shakey

Shakey was developed in the Stanford Research Institute in 1966-72. It was the first mobile robot able to reason about its actions. Shakey had a TV camera and a triangulating range finder in a movable head, and bump sensors. It had two stepping motors driving its wheels and other motors to control the camera focus and the tilt angle of the head. It was connected to DEC PDP-10 and PDP-15 computers via radio and video links (Nilsson, 1984).

Shakey could move on flat surfaces between office rooms and perform simple tasks that required planning, route-finding and rearranging of simple objects. It maintained an axiomatic model of its world (see Figure 4.1) and used a hierarchy of programs for perception (including simple vision), world-modeling, and acting. Low-level routines, represented by symbolic commands with parameters, e.g. `TILT numberOfDegrees`, provided the interface between the robot's hardware and higher-level software and took care of simple moving, turning, panning and taking a TV picture. Axiomatic predicates of this level represented the robot's low-level state, e.g. `(AT ROBOT xfeet yfeet)` or `(TILT ROBOT degreesUp)`. Intermediate level actions were described in terms of going to a specified place or pushing an object from one place to another, e.g. `GOTHRUDR(DOOR FRROMRM TORM)` moves the robot from room `FROMRM` to room `TORM` via door `DOOR`.

On the highest level, Shakey used a planning system STRIPS[3] (Fikes and

[3]Acronym for STanford Research Institute Problem Solver.

Operator:
GOTHRU(d, r1, r2)
Preconditions:
INROOM(ROBOT, r1) ∧ CONNECTS(d, r1, r2)
Delete List:
INROOM(ROBOT, $)
Add List:
INROOM(ROBOT, r2)

Figure 4.2: An example of operator specification in STRIPS. Strings beginning with lower-case letters are parameters, $ denotes an arbitrary value. The example is taken from Nilsson (1984).

Nilsson, 1971) that could chain together intermediate-level actions into plans and execute them to achieve goals given by a user. The goals were specified as first-order predicate logic formulas.[4] In order to plan a sequence of actions that would change the world in such a way that the current goal formula is true in the changed world model, STRIPS needed to know the effects of each action. The model of each action, called the *operator*, consisted of *preconditions* (formulas that were required to be true in the world model for the operator to be applicable) and *postconditions* in the form of an *add list* and a *delete list*. The effect of applying the operator to a world model consisted in deleting from the model all those clauses specified by the delete list and adding to the model all those clauses specified by the add list (see Figure 4.2).

STRIPS also had some learning abilities – it was able to generalize plans by using parametrized operators and by identifying subsequences of actions necessary for the success of the new plan. It saved these generalized plans for possible future use.

Evaluation. Shakey was a physically realized robot, situated in the real (albeit simplified) environment, that was interacting with the world via its sensors and motors. Its (mostly preprogrammed) knowledge about the world was represented by symbolic first-order predicate logic formulas. It "understood" the world around to the extent that it possessed the representation of consequences of its actions and used it to plan actions in order to achieve goals by modifying the environment.

[4] Actually, users gave instructions in simplified English, e.g. "USE BOX 2 TO BLOCK DOOR DPDPCLK FROM ROOM RCLK." and the statements were subsequently converted by the system ENGROB (Coles, 1969) to first-order logic formulas, e.g. BLOCKED(DPDPCLK,RCLK,BOX2).

4.2.2 ILM: Modeling the Emergence of Grammar

The *iterated learning model* (ILM) framework (Kirby and Hurford, 2001) has been designed for modeling the emergence of compositional structures (grammar) in a language by means of cultural transmission. The model usually consists of two agents – an adult teacher and an infant learner. The teacher uses its language competence[5] to generate a linguistic input for the learner in the form of utterance/meaning pairs. The learner, initially having no knowledge of the target language, uses this input to induce its own linguistic competence. Later, the original teacher is removed from the population, the learner becomes a teacher for the next-generation learner and the whole process iterates. The key result of the ILM models is that, in the process of iterated cultural transmission, language itself undergoes changes and adapts for better transmission by incorporating structural regularities.

There are many instantiations of ILM framework; here we review one of the experiments that use predicate logic for representation of meanings. In the experiment of Kirby (2000), meanings were drawn from a predefined meaning pool common to all agents. Meanings were compositions of symbolic atoms divided into two classes:

$$Objects = \{Mike, John, Mary, Tunde, Zoltan\}$$

and

$$Actions = \{Loves, Knows, Hates, Likes, Finds\}\,.$$

Each meaning was a triple $\langle Agent, Patient, Predicate\rangle$, wherein only atoms of the type *Object* could stand for the *Agent* and the *Patient*, and only atoms of the type *Action* could stand for the *Predicate*. Within these type restrictions, the atoms could be combined to form the total of 100 meanings.[6] In a more readable fashion, the meanings could be written as predicate-argument propositions $Predicate(Agent, Patient)$.[7]

The meaning space was given beforehand and did not change during the simulation. In the production phase of an iteration, the adult agent was given a random subset of the meaning space and had to produce utterance/meaning pairs for each meaning in the subset. Utterances were generated by the adult agent's grammar that embodied its linguistic competence. The production

[5]The first-generation teacher has no linguistic competence either. However, it uses an invention mechanism and generates random strings for parts of a sentence that it otherwise could not express.

[6]Reflexive meanings with the same atom substituted for both the *Agent* and *Patient* were not allowed.

[7]In Kirby (2002), meanings were extended to include recursive structures such as *knows(john,eats(tiger,john))*.

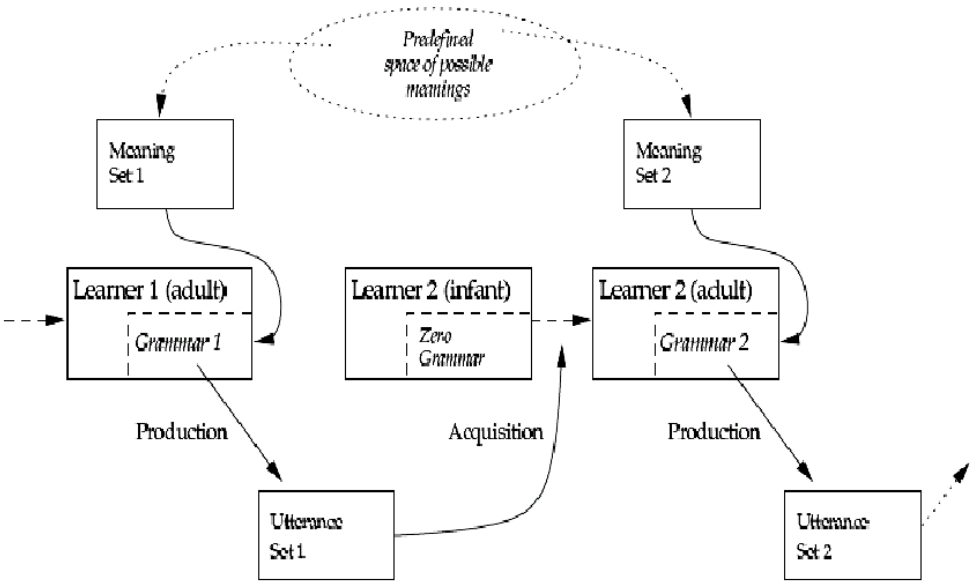

Figure 4.3: The Iterated Learning Model (ILM). The adult agent uses its grammar to produce utterances for a randomly chosen subset of meaning space. The produced utterance/meaning pairs are used by the infant agent for grammar acquisition. After inducing a grammar, the learner becomes an adult and produces a teaching input for the next-generation infant. The whole process then iterates. The picture is taken from Kirby (2002).

included random invention in case the agent had no rule to generate a string for some meaning or its part.

The utterance/meaning pairs served as the input for the induction of the learner's grammar[8] (see Figure 4.3). The agents used a simple version of a context-free definite clause grammar (DCG) enriched with statistical information and semantic components. The grammar could encode holistic rules as well as compositional ones (see Figure 4.4). After receiving a new utterance/meaning pair, the agent stored it in the form of a holistic rule. Holistic rules could later be subsumed (generalized) to compositional rules by utilizing structural regularities that randomly occurred in the utterances.

Evaluation. In the reviewed experiment, the meanings were represented by symbolic predicate-argument propositions. However, meanings were external to the agents, given beforehand, fixed and common to all the agents. They bore no relation to any environment; in fact, there was no environment at all in this experiment. The agent's task was to learn a mapping between

[8]Each agent initially started with an empty grammar.

A holistic rule:
S/*eats(tiger, john)* → `tigereatsjohn`

Compositional rules:
S/*p(x,y)* → N/*x* V/*p* N/*y*
V/*eats* → `eats`
N/*tiger* → `tiger`
N/*john* → `john`

Figure 4.4: Holistic and compositional portions of an example DCG grammar used in ILM. Capital letters denote non-terminals and typewriter-style strings denote terminals (substrings of a generated utterance). Lower-case italic variables denote preterminals that can be substituted with semantic components (parts of meanings, which are represented by strings after the slashes and written in italic). The example is taken from Kirby (2002).

a static compositional code that the experimenters called meanings, and a new emerging compositional system called syntax.

Another unrealistic assumption of this model was that, together with an utterance, the learner received its meaning. This "telepathic" approach created the *signal redundancy paradox* (Smith, 2003a): If the meanings are directly transferable, then the signals are redundant.

4.3 Uninterpreted Scalars and Vectors

Omitting the environment and abstracting away semantic problems has been quite common in computational models that primarily view the language acquisition/evolution as the problem of acquisition/creation of correct mappings between signals and meanings.

4.3.1 Formal Models of Innate and Learned Communication

In the formal models of Oliphant (1997, 1999), communication is analyzed in terms of relations between *meanings* and *signals*, where each meaning μ is represented by an integer, e.g. 1, 2, 3, and each signal σ by a letter, e.g. a, b, c.[9] The communicative behavior of an individual is specified by

[9] A meaning is said to represent an associated pair ϵ, α of some environmental state ϵ and an action α appropriate in the state ϵ, however the pair is never used in the course of

two probability functions: *transmission* function s and *reception* function r, where $s(\mu, \sigma)$ represents the probability that the signal σ will be sent for a meaning μ, and $r(\sigma, \mu)$ represents the probability that a signal σ will be interpreted as the meaning μ by the individual. Communication between a sender and a receiver is successful, if the meaning of the sender's signal, as decoded by the receiver, is the same as the speaker's intended meaning. A measure of *communicative accuracy* is defined as an average probability that any given meaning will be successfully communicated (between two individuals, or in the whole population).

In Oliphant's experiments, transmission and reception functions of individuals are encoded either in genome matrices in the models of innate (genetically evolved) communicative behavior, or in various types of simple networks (Willshaw networks, Cummulative-Association Networks and Hebbian networks)[10] in the models of learned communicative behavior. The experiments evaluate various learning strategies in terms of achieved communicative accuracy.

Evaluation. The models bring valuable theoretical insight into the problem of the acquisition of correct associations between meanings and signals. However, the problem of meaning formation is abstracted away: meanings are unstructured tokens (integers) given in advance, fixed and common for all the agents, without a relation to any environment.

4.3.2 Emergence of Syntax

From the models of emergence of syntax that use uninterpreted binary vectors as meanings, we have chosen the model of Kvasnička and Pospíchal (1999). The purpose of this model is to study a hypothesis that coordinated communication together with grammar regularities are results of an evolutionary process running in a population of agents. Cognitive devices of agents are represented by neural networks, similarly to the model of Batali (1998). The difference between the reviewed model and that of Batali is that while the latter model studies the emergence of structural regularities within one generation of neural networks, the former model is enlarged with Baldwin effect (Baldwin, 1896) and Dawkins memes (Dawkins, 1976) within the evolutionary process spanned over many generations.

Let us look at the model of Kvasnička and Pospíchal (1999) in more detail now. The experiments are performed with the population of 30 agents. Each agent uses a simple neural network with one layer of hidden neurons for

experiments, nor there is any environment.

[10]For details, refer to the original source.

coding meanings into signals and vice versa. Meanings are 8-dimensional binary vectors $\vec{y} \in \{0,1\}^8$ that model internal states of the agents. Meanings have an internal structure; the set of allowed meaning vectors is the Cartesian product[11]

$$A = \{(1100), (0011), (0101), (1010)\} \times \{(0110), (1001), (1100), (0011)\} \ .$$

The meaning vectors are coded by speakers to strings of maximal length 10 composed of two symbols {a,b}.

Each neural network can be specified by its structure, connection weights and thresholds. Such specification is used as the agent's genome in the evolutionary process. This deliberately vague specification allows to represent several types of neural networks ranging from feed-forward to recurrent ones. Hence, the population can be composed of agents equipped by neural networks with different architectures.

The evaluation of the fitness of an agent is based on its success in pairwise communications with other agents (the success is inversely proportional to the difference between speaker's meaning and the meaning decoded by the listener). The set of allowed meaning vectors A is divided into a training set and a testing set. For all possible pairs speaker – listener all meaning vectors from the training set are applied in elementary communication acts. During the communication acts, the listener's neural network is modified so that the difference between the speaker's and listener's meaning vectors decreases. At the end of the generation, a new population is created by quasi-random selection of best-fitted networks that undergo small mutation. The offspring inherits from parents only their initial weights and threshold coefficients, in order to prevent Lamarckian inheritance.

To study the effect of Dawkins memes, each agent generates a training set of "memes" in the form of meaning/signal pairs in the end of its life. This set is inherited by the agent's offspring, which is trained on it before entering the communication acts.

The results of simulations have shown that:

1. in the course of evolution, the decoded meanings gradually get closer to meaning vectors of speakers and all the agents gradually start to use the same vocabulary for the common communication,

2. regularities in meaning vectors are manifested also in the structure of signals (similar parts of meaning vectors are coded by similar symbol

[11]Batali (1998) suggests that such vectors can be interpreted as predicate-argument clauses. This interpretation is not used in the reviewed model.

substrings); this is considered a manifestation of an emergence of a grammar system,

3. for the emergence of coordinated communication, the inclusion of memes is of primary importance.

Evaluation. This model is an important step toward understanding the roles of Darwinian evolution and cultural transmission of memes in the emergence of syntax. However, with respect to our criteria put on meanings, the model contains many simplifications. First, there is no environment in the model. The meanings are external to the agents, predefined and fixed during the whole experiment. Second, the model relies on the unrealistic assumption that the listener has a direct access to the speaker's internal state (the meaning). It is used both for training from the meme set and for adapting the listener's network during the communication acts. Although the authors presume that the listener can find out the internal state of the speaker because it corresponds to some external surrounding reality which can be determined also by the listener, the inference of meaning from a situation is not easy and is subject to referential indeterminacy or Gavagai problem (see Section 3.3).

4.4 Regions in a Space

In this section, we describe several models that consist of agents perceiving a shared environment through their sensors. A common feature of these models is that meanings are represented by regions in some geometrical space, typically a space defined over possible sensor values. Moreover, meanings are not externally given in these models, but created individually by each agent.

The agents are either software entities in a simulated environment, or programs embodied in real robots operating in a real environment. Each agent senses its environment through a set of sensors that give readings $\{s_1, \ldots, s_n\}$ with ranges $s_i \in D_i$ (in most of the experiments, normed ranges $D_i = (0, 1)$ are used for all sensors).[12] Meanings then typically represent classes of sensor readings that should be treated equally with respect to some purpose. These classes form regions either in single-sensor 1-dimensional spaces D_i, or in the multidimensional space $D_1 \times \ldots \times D_n$ in case that tuples of all sensor values are considered simultaneously.

[12] A tuple of sensor readings can represent the whole perceived scene that possibly consists of several objects. However, in many models, sensor readings are preprocessed in that a separate n-tuple is given for each object in the agent's vicinity.

4.4.1 Games the Agents Play

The agents enter into mutual interactions called *language games* that proceed in rounds. In each round, one agent is selected as a *speaker* and another as a *hearer*. A typical communicative goal of the speaker is to uniquely identify a particular object present in the environment, chosen as the *topic*. All other objects concurrently present in the environment form the *context*.

The speaker first plays a *discrimination game* (Steels and Kaplan, 1999) in order to find among its internal meanings one that distinguishes the topic from all other objects in the context. In case of failure, the representation of meanings is refined. In case of success, the speaker tries to lexicalize the selected meanings.

The lexicon of an agent typically consists of many-to-many associations between words (strings of characters) and meanings. Each association has a strength expressed by a positive real number. The agent learns associations between words and meanings by manipulating the strengths of the associations based on success/failure in language games or on noticing word-meaning co-occurrences. There are several variants of how this can happen:

- In the *guessing game* (Steels and Kaplan, 1999), the speaker utters an expression and the hearer tries to guess what referent the speaker names. Afterwards, the hearer receives a feedback indicating whether its guess was correct or not.

- In the *observational game* (Oliphant, 1997), the speaker narrows down the set of possible referents by e.g. pointing, and the hearer adapts its lexicon by Hebbian learning. In case there are still more than one candidate left, the hearer associates the speaker's verbal description with each of the candidates (the hearer receives no feedback in this case). Meanings can then be disambiguated cross-situationally (Siskind, 1996; Smith, 2005a).

In all the models, the hearer's inference of the meaning is constrained by the assumption that the scene only contains one referent of the speaker's utterance (semantic hypotheses with more referents are excluded from consideration). In some models, the hearer also assumes mutual exclusivity, i.e. it excludes from consideration all those objects on the scene, for which it already knows an appropriate word (Smith, 2005b).

Now we will review several possibilities of representation of meanings by regions in the sensor space in more detail.

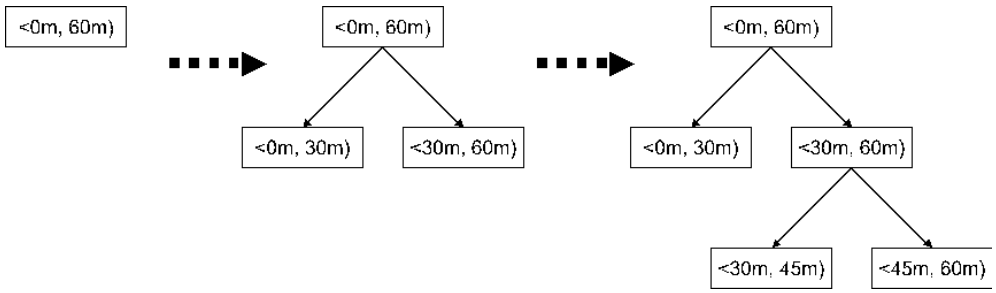

Figure 4.5: Splitting a discrimination tree. The range of the sensor is repeatedly split into halves. Note that the tree does not have to be balanced. The resulting density of the splits depends on a particular history of discrimination games in particular environmental contexts. The picture is taken from Bodík and Takáč (2003), where we modeled the emergence of a common spatial lexicon in agents traveling over the lattice.

4.4.2 Discrimination Trees

One possibility of partitioning the sensor spaces is via *discrimination trees* (Steels, 1997), which are used in models of language formation based on computer simulations (Bodík and Takáč, 2003; Smith, 2003a, 2005a) or experiments with real robots (de Jong and Vogt, 1998; Steels and Vogt, 1997; Steels and Kaplan, 1999; Steels, 1999; Steels et al., 2002). For each sensory channel, the agent constructs a separate binary discrimination tree. Nodes of the tree represent subintervals of the corresponding sensor's range. They determine the granularity of the agent's representation: all sensor readings that fall within an interval of some node are treated equal.

Initially, each tree only consists of a root that represents the whole range of the corresponding sensor, e.g. the interval $[0, 1]$. A node can spawn two children that represent halves of their parent's interval, e.g. $[0, 0.5)$ and $[0.5, 1]$. Upon failure in a discrimination game, agent randomly chooses a node of some tree and splits its interval into halves by spawning two subnodes (see Figure 4.5). The utility of each split is monitored by recording its use and success in future discriminations. Environmentally irrelevant (unused or unsuccessful) distinctions will be discarded.

We will illustrate this in more detail on a paradigmatic experiment *Talking Heads* (Steels, 1999). The experimental setup consists of two pan-tilt cameras in which different agents can be loaded. The agents loaded into the cameras perceive a shared environment that consists of a magnetic white board on which various geometric shapes of various colors are pasted: triangles, circles, rectangles, etc. Agents are capable of segmenting the perceived image into

objects and of collecting various characteristics about each object, specifically the color (decomposed in RGB channels), grayscale, and position in pan/tilt coordinates. Hence, each object on the scene is represented by a n-tuple of (preprocessed) sensory values.

In this experiment, agents play guessing games. A speaker first selects a topic and plays a discrimination game. It searches through each discrimination tree and tries to find a node with the interval that contains the corresponding sensor's reading of only the topic (and no other object from the context). If it cannot uniquely identify the topic by a node from a single discrimination tree, it tries to combine nodes from several trees. In case of success, the speaker verbalizes each of these meanings (each describing some feature of the topic) by finding words associated with these meanings in the lexicon. If a meaning is associated with several words, it can simply choose the strongest association, or use the *introspective obverter* strategy (Smith, 2003b) by selecting the word that the speaker itself would best understand as the meaning (if it does not have such a word, it can invent a random one).

The hearer then tries to decode the uttered words by inspecting associations stored in its own lexicon. After finding the associated meanings, it tries to use them to uniquely identify some object on the scene, i.e. to guess the object meant by the speaker. The speaker and hearer receive feedback about the result of the game and adapt scores of their representations accordingly. In case of failure, the speaker identifies the intended topic to the hearer by pointing. Note that the evaluation of success in the game is based on a pragmatic criterion of *referent* identity: the agents must agree upon an external object, but each of them can use different internal meanings to represent the object.

The experiment ran for 4 months in 1999. There were close to 6000 agents launched and they played 400000 grounded language games. The agents started with empty lexicons and meaning repertoires. A shared lexicon enabling successful communication emerged within a few days. A total of 8000 words and 500 concepts were created, with a core vocabulary consisting of 100 basic words (Steels et al., 2002).

Evaluation. The discussed experiment Talking Heads is a practical demonstration of the emergence of a common lexicon from scratch in a community of physically realized agents by means of purely horizontal interactions within one generation of agents. The experiment explores the role of cultural transmission in language evolution, as there is no genetic selection in the model.

Each agent creates its meanings individually based on the interaction with the real environment. The meanings are private, not available to other agents and possibly different in each agent. The agents communicate about static

objects currently present in their surrounding. Although the agents possess multiple sensory channels, they do not relate them multi-dimensionally. Hence, meanings (nodes of trees) do not actually represent objects, but abstractions over feature values. They can be situationally combined to refer to objects, but a persistent representation encoding covariant properties is not feasible in this model. Categories (intervals of feature values) have sharp boundaries. They are mutually exclusive (non-overlapping) on the level of leaf nodes of the tree – a particular object's feature value can fall into just one interval. The whole tree supports hierarchical relations (an interval of a node higher in the tree includes intervals of its subtree). As the lexicon contains many-to-many word/meaning associations, it supports synonymy and homonymy.

Distinctions expressed by a discrimination tree are created situationally and kept environmentally relevant (by pruning irrelevant branches). Actual shape of the representation is optimized to the goal of success in discrimination games. Whether such a driving force leads to cognitively plausible categories, is an open question. We discuss this issue in Section 4.4.5.

4.4.3 Situation Concepts

A *situation concept* is a subset of the possible histories of an agent's interaction with its environment with the property that knowing to which situation concept the actual history of interaction corresponds, allows the agent to predict some aspect of the future (de Jong, 1999). Each agent constructs situation concepts individually by observing patterns in the sequence of inputs from the environment, its own actions, and subsequent evaluative[13] feedback. Situation concepts were used in the model of de Jong (2000) inspired by the alarm call system of vervet monkeys described by Seyfarth et al. (1980). The innate[14] system of alarm calls enables vervet monkeys to distinguish three types of predators: birds of prey, large mammals and snakes. Based on signals from other monkeys, a monkey can take the appropriate flight action even if it did not detected the predator directly itself.

In de Jong's experiments, five agents can move horizontally and vertically on a grid. Each agent has three sensors: one (S_1) indicating the type of a predator that is present (or the absence of predators), two (S_2 and S_3) for the

[13] *Evaluative* feedback is available in the form of a numeric reward corresponding to the appropriateness of the performed action. It should be distinguished from *instructive* feedback, which consist in providing direct information about the appropriate (or required) action. In the experiments described in this section, only evaluative feedback is at the agents' disposal.

[14] In the model, the set of signals and their meaning is learned.

agent's own horizontal and vertical coordinates. Actions consist of moving horizontally one step left or right or staying (action dimension A_1), and selecting a vertical position (action dimension A_2). A predator of random type (three different types of predators exist) is created in 10% of the time steps at a random horizontal position, provided no predator is present yet. The vertical position of an agent determines whether it is safe from the predator or not, and since the number of vertical positions is three, each position corresponds to a single type of predator. The horizontal position of an agent determines whether it can see the predator. The scope of the agents' perception amounts to 90% of the field; hence, for each agent, 10% of the predators are expected to be invisible (de Jong, 1999).

If, during the presence of a predator, an agent is not in the safe row, it will receive a zero reward (evaluative feedback). If it moves to the safe row, it will be rewarded by 1.0. When no predator is present, staying at the same place is rewarded by 1.0 and all the other choices of vertical movement are rewarded by 0.5. Each agent starts with a 5-dimensional sensor-action space, where each dimension corresponds to the whole range of possible sensor or action values. During the experiment, each agent adaptively splits the whole space into halves, depending on whether a split along a dimension leads to a more uniform distribution of rewards between the two halves. The history of all splits is represented by a tree (see Figure 4.6). Not all the sensor dimensions are relevant for choosing the appropriate action: only the sensor dimension indicating the predator type should correlate with the chosen vertical position; all other dimensions are irrelevant. The agents should discover this by utilizing the history of interactions and rewards.

In each time step, an agent receives signals from all the other agents indicating the type of the perceived predator. Bayesian conditional probability rule is then used by each agent to decide whether to rely on the observed type of the predator, or the predator type decoded from the signals uttered by the other agents.

Evaluation. De Jong's experiments are conducted in a simulated environment. The meanings represented by subregions in the state-action space are individually created by each agent, based on the external pragmatic evaluative feedback. The concepts integrate situation and action categories in that they support causal predictions of the effect of actions in a particular situation (in terms of the expected yielded reward). All categories have sharp boundaries (i.e., each input is categorized in one and only one category). The representation supports hierarchic meanings in terms of nodes on different levels of the tree. Unlike in the discrimination trees (see the previous section), categories are formed as multidimensional regions in the state-action space. Hence, they implicitly encode correlations between dimensions. The

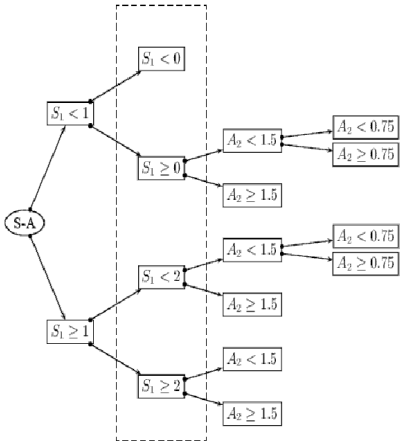

Figure 4.6: Situation concepts: De Jong's representation of meanings in a five-dimensional hyperspace using adaptive splitting to subspaces. Distinctions have been made in the S_1 and A_2 dimensions corresponding to a perceived predator's type and a move in the vertical direction. The picture is taken from (Smith, 2003a).

driving force of the meaning formation process is the feedback about success in the predefined task (to choose a particular action corresponding to the type of the present predator). This learning mechanism has been criticized (Smith, 2003a): failure in accomplishing the task could be fatal – the first time an agent chooses a wrong action, it would be caught or killed. Hence, the environmental feedback would not allow the agents to learn and to solve the problem.

4.4.4 Prototypes

The general idea of prototypes comes from the empirical findings that some exemplars of categories are more representative than others (Rosch, 1978). In computational models, prototype representation allows for better cognitive economy: only the best exemplars (prototypes) need to be remembered. Category membership of any input is then determined by its distance from prototypes in some geometric *conceptual space* (Gärdenfors, 2000, see Section 2.4.5): each point of this space is considered a member of the category represented by the spatially closest prototype (see Figure 2.1). Prototypes were used as representation of meanings in experiments with software agents in a simulated environment (Vogt, 2003a,b, 2005; Divina and Vogt, 2006; Vogt and Divina, 2007) as well as with real robots (Vogt, 2000, 2002). We will illustrate the usage of prototypes in more detail on the experiment con-

ducted by Vogt (2002). In this experiment, language games are played by two agents embodied in two mobile LOGO robots. The goal of a language game is to communicate a name for one of light sources that the robots can detect in their environment. Both robots first sense their surroundings, then one of the robots takes the role of the speaker and the other takes the role of the hearer. The speaker selects one sensation of a light source as the *topic* and individually plays a discrimination game to find a unique category that distinguishes the topic from the context – sensations of other present light sources. Then it searches a word that it has associated with this category in the past, and communicates this word to the hearer. The hearer, who has also sensed several light sources, tries to interpret the communicated word. The language game is successful when both robots communicated about the same light source; in this case, the agents increase scores of the structures and associations used in this game. If the language game was not a success, the lexicon has to be adapted either by creating a new form (if the speaker could not produce an utterance), by adopting the form (if the hearer could not understand the utterance) or by decreasing association scores. In the beginning of the experiments, the robots have no categories or words; these are developed during the language games.

In the sensing phase, a stream of raw sensory data is preprocessed, so that all perceived light sources can be represented by feature vectors of the same length. Hence, the context of m perceived light sources $\{S_0, \ldots, S_m\}$ can be represented as m real-valued vectors/*points* $\left\{\vec{f}_1, \ldots, \vec{f}_m\right\}$ in a n-dimensional feature space $\mathcal{F}$. Categories are *regions* in the feature space. Each point of the feature space belongs to a category defined by the closest prototype. Each agent maintains several versions $\{\mathcal{F}_\lambda \mid \lambda = 0, 1, \ldots, \lambda_{max}\}$ of the feature space with different resolution of at most 3^λ values on each dimension.

During the discrimination game, the agent categorizes the topic by finding the set of closest prototypes $\{\vec{c}_0, \ldots, \vec{c}_{\lambda_{max}}\}$ in all versions $\mathcal{F}_\lambda$ of the feature space $\mathcal{F}$. If no other light source from the context falls in the same category set, the discrimination game is successful. If the category is used as the meaning in a language game successfully, its prototype is shifted toward the feature vector of the topic, so that the prototypical category becomes a more representative sample of the feature vector it categorized. If the game fails, some feature f_i of the topic is randomly selected and a feature space $\mathcal{F}_\lambda$ that has less then 3^λ values on the i-th dimension is refined by adding new prototypes $\vec{c}_j = (x_0, \ldots, x_{n-1})$ where $x_i = f_i$ and the other x_r are made of already existing prototypes in $\mathcal{F}_\lambda$. All prototypes have several scores associated, which are used during the selection process in language games.

Evaluation. Vogt's experiments show how a coordinated grounded lexi-

con can emerge in physically realized robots that communicate in real conditions. Meanings are constructed individually by each agent from real sensory data to represent currently present objects (light sources). More complex meanings (relations, situations, events) are not feasible in this model. The model supports synonyms and homonyms. The existence of multiple versions of the feature space can be seen as a support for hierarchical relations. Categories within one version of the feature space have sharp boundaries. Due to the prototype addition mechanism that has the same effect as splitting the feature space in half, categories have hyper-rectangular rather than hypher-spheric shapes and correspond to multidimensional discrimination trees. Some feature distinctions inherited due to this mechanism may be unnecessary for some categories. Truly multidimensional categories can hardly be represented in other way then by a bag of neighboring prototypes. Success in discrimination games is the driving force of meaning creation.[15]

4.4.5 Discrimination Versus Identification

In most of the just-described models, the communicative goal is to uniquely identify a chosen static object by discriminating it from all other objects currently present in the communicative situation. The proposed representation-forming mechanisms are tailored to this goal; they are based on capturing differences between the present objects.

The communicative goal of unique discrimination in a particular context determines the shape of the meanings. The categories that have evolved for the purpose of discrimination do not have to be natural and suitable for other purposes, as has been argued by Harnad (2005), who distinguishes between discrimination and identification (categorization). Discrimination is a *relative* judgment between things that are present simultaneously, while identification (categorization) is an *absolute* judgment of a thing alone answering the question whether or not a given input is a member of a particular category (Harnad, 1990, 2005).

Language goes beyond a present situation, and its important function is to enable detached communication about things not present here and now (Gärdenfors, 1996b). For detached use of language, the importance of identifying categories (kinds) of things, which is apparently a different situation from a discrimination task, is even higher. We argue that, for this goal, rep-

[15]A noteworthy exception is the New Ties project (Divina and Vogt, 2006; Vogt and Divina, 2007) – a simulation of large-scale (more than 1000) agent population that evolves using a combination of evolutionary, individual and social learning. In this experiment, agents sense, act, eat to regain energy, talk and mate, and die if they get too old or run out of energy.

resentation based on noticing *similarities* rather than differences is more suitable (see Section 3.3.2). More importantly, meanings should not be limited to object categories; they should also include properties, relations, dynamic changes, situations and events.

4.5 Meanings in Dynamic World

In this section, we broaden our perspective on meanings by models that go beyond static objects and are capable of representing dynamic aspects of the world, which is a necessary prerequisite for verb understanding.

4.5.1 Redescriptions of Co-Occurring Events

The model of Cohen et al. (1996) focuses on how adult concepts can develop from infant sensorimotor activity. It consists of an artificial baby – Neo living in a simulated environment. The environment implements Neo's sensations, mental and physical activities and the behavior of other objects and agents that interact with Neo.

Neo senses its environment through a collection of *streams*, which can hold different symbolic *tokens* in different discrete time steps. For example, one token is **rattle-shape** and it is placed in the appropriate stream whenever Neo's eyes point at an object that is shaped like a rattle. The streams that represent Neo's internal sensations include an affect stream that contains tokens such as happy and sad, a pain stream, a hunger stream, and somatic and haptic streams that are active when Neo moves and grasps. Neo performs random actions: it can move its arm and head, and grasp several objects, including three rattles, a bottle, a mobile, keys and a knife. The latter causes pain. The rattles make noise when shaken. Neo gets hungry some time after eating, it cries when it is unhappy or in pain; when Neo cries, Mommy usually visits, unless she is angry at Neo for crying, in which case she stays away.

Neo can learn representations of objects, states and activities by using several versions of a simple learning rule based on noticing temporal regularities, i.e. co-occurrences of values in streams or in representational structures (*redescriptions*) built on top of streams. Neo develops and utilizes five kinds of redescriptions of its sensations:

1. *Changes* in token values: Neo notices time steps when a stream changes its value and also time steps when no change happens.

2. *Scopes*: By maintaining contingency tables for pairs of streams, Neo finds correlated streams that change together often – the scopes. Scopes

provide a mechanism for cross-modal perception.

3. *Base fluents*: Neo finds time intervals with co-occurring tokens within scopes, e.g. the base fluent **((sight-color red) (sight-shape rattle-like))** that represent a red rattle and **((sight-color dark) (sight-shape none))** that represents what happens when Neo closes its eyes.

4. *Context fluents*: Neo finds base fluents that tend to follow each other in time, e.g. **(CONTEXT ((sound cry) (mouth not-mouthing)) ((tactile-hand plastic) (hand close)))** represents an experience of crying Neo who grasped a plastic object.

5. *Chains*: These temporal dependencies are combined into temporal chains, which represent activities, e.g. **(CHAIN ((tactile-mouth none) (voice cry)) ((tactile-hand wood) (hand close)) ((tactile-mouth wood) (do-mouth mouth)))** is a representation of crying Neo as it grasps and mouths a wooden object. Chains are used for activity-based categorization.

Evaluation. This experiment has demonstrated that conceptual structures such as image schemas (Lakoff, 1987) do not have to be innate and can develop by simple temporal associative learning from streams of sensorimotor data. Neo lives in a simulated world that consists of streams of symbolic tokens. Yet it autonomously constructs basic meanings by observing temporal regularities in its sensations. Albeit on a very elementary level, meanings are not limited to static objects, but include representations of changes, interrelated processes and chains of events. Such sensorimotor-based concepts suggest how a language can bootstrap from a preverbal stage and provide grounding for more abstract meanings via metaphoric mappings (Lakoff and Johnson, 1980).

4.5.2 Dynamic Maps (Phase Portraits)

The work of Cohen (1998) focuses on representation of verb meanings based on *dynamic maps* that capture the dynamics of interactions between two agents or objects. The maps are constructed for three phases of interaction: before, during and after contact. Each map is a two-dimensional qualitative[16] phase portrait. Trajectories in these phase portraits correspond to different types of interactions and are denoted by different verbs.

[16]In the sense that continuous real axes are abstracted to regions of negative, zero and positive signs.

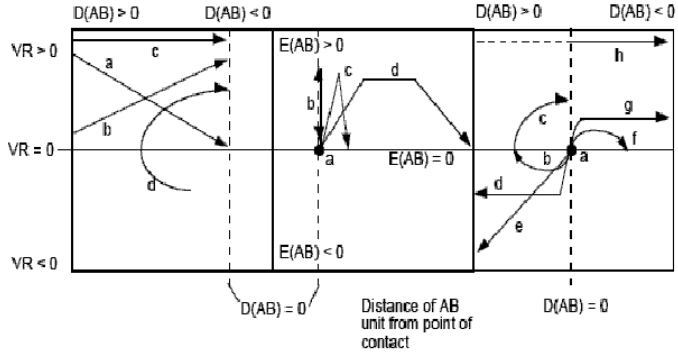

Figure 4.7: The before, during and after phases of physical interactions between A and B. The dashed vertical lines represent the point of contact, $D(AB) = 0$. In the before and after phases, regions to the left of $D(AB) = 0$ represent A behind B and regions to the right represent A ahead of B. In the during phase, regions to the right of $D(AB) = 0$ represent displacement of the AB unit (remaining in contact) from the point of contact. Combination of trajectories allows for representation of number of verbs: A gently touches B (*aaa*), A pushes B (*ada*), A kicks, propels, shoves, bounces off B (*adb, acb*), A leans/strains against B (*aca, ada*), A dislodges, frees B, or B flees from A (*bce*), A crashes into B (*cba*), A hammers, harasses, claps B (*dcc*), A pushes through B (*bbf*), A breaks free of B (*bbg*). The picture is taken from Cohen (1998).

Dimensions of the maps for the *before* and *after* phase of the interaction are the relative position or distance $D(AB) = P(A) - P(B)$ between the two bodies (actor A and target B) and their relative velocity $VR = V(A) - V(B)$, where $P(A)$ and $P(B)$ are projections of (not necessarily physical) locations of A and B onto a one-dimensional *progress space*,[17] $V(A) = \mathrm{d}P(A)/\mathrm{d}t$ and $V(B) = \mathrm{d}P(B)/\mathrm{d}t$. The map for *during* phase has the distance of the connected bodies from the point of contact on the horizontal axis and the energy transfer between the bodies (acceleration of the actor in the direction of target, while they are contact) on the vertical one (by definition, the distance between A and B is zero and so must be their relative velocity, otherwise the distance would change). Combination of trajectories from phase portraits for before, during and after phases allows for representation of number of verbs, see Figure 4.7.

Evaluation. The project focuses on the representation of the dynam-

[17]This allows for representation of non-physical verbs that only describe movement metaphorically, e.g. verbs for transfer or advancement of wealth, information, skills or credibility.

ics of the world. Dynamic maps can compactly and explicitly represent the manner of movement, which is very important for capturing subtle distinctions between verb meanings. They can be easily learned and recognized from sensory-observable information. Preliminary experiments with supervised learning of dynamic maps and their usage for predictions are reported in Rosenstein et al. (1997). The learning is supervised in the sense that the system is told the class of behavior it is observing, and it merely learns the dynamics of the interaction. The authors have also developed an unsupervised version, where the system clusters training trajectories together without knowing which behaviors generated them (Cohen, 1998).

The proposed representation can also be useful for object conceptualization, as the classes of objects are differentiated by the way we interact with them (Gibson, 1979; Lakoff, 1987). Again, we remark that dynamic maps are not limited to verbs denoting physical action: meaning of many abstract verbs can be metaphorically described in physical terms (Lakoff and Johnson, 1980).

4.5.3 Semiotic Schemas

A unified representational framework for meanings of verbs, adjectives and nouns was proposed by Roy (2005b). In this framework, all meanings are represented by *schemas*. Schemas are constructed in bottom-up fashion from *beliefs* by six types of projections (sensors, actions, transformers, categorizers, intentional projections, and generators). Beliefs (analog or categorical) are probability distributions over *analog signs* (patterns of sensor readings) or *categorical signs* (discretizations of analog signs).

Schemas encode the knowledge necessary for interpreting, verifying, and guiding actions towards objects, object properties, spatiotemporal relations, situations, and events: Objects are represented by networks of interdependent schemas that encode properties and affordances, verbs are grounded in sensory-motor control programs, adjectives describing object properties are grounded in sensory expectations relative to specific actions. Locations are encoded in terms of body-relative coordinates. Speech acts of agents are interpreted either into goal schemas that an agent may choose to pursue, or into existential beliefs represented through schemas which are compatible with sensing and action. For example, the meaning of "red" is a color category linked to the motor program for directing active gaze toward an object, and "heavy" is grounded in haptic expectations associated with lifting actions. The meaning of "ball" subsumes both the meaning of "round" (which is one of its expected properties along with color, size, etc.), and all of the actions that may affect the ball (Roy, 2005a).

Semiotic schemas were implemented in a manipulator robot Ripley designed for grounded language experiments (Roy et al., 2004). Ripley has a gripper with seven degrees of freedom driven by actuators instrumented with position and force sensors, providing the robot with a sense of proprioception, and two miniature video cameras. Ripley's gripper fingers are instrumented with force-resistive sensors giving it a sense of touch. The robot's work space consists of a round table. Its motor control system allows it to move around above the table and view the contents of the table from a range of visual perspectives. Several other motion routines enable the robot to retract to a home position, to lift objects from the table, and to drop them back onto the table. Ripley is able to translate spoken commands such as "hand me the blue one on your right" into situated action.

Evaluation. Semiotic schemas framework is a complex and comprehensive step toward language grounded in the real world. Complex meanings of verbs, objects, properties and events are represented in a unified fashion that enable prediction, inferences, interpretation and generation of behavior and planning. Being implemented in a real robot, this approach does not abstract away from issues of motor control, proprioception, cross-modal coordination, etc. In this way, meanings depend on the way the robot is constructed and are truly embodied. An unresolved issue is the origin of schemas: although the robot can determine settings of some parameters by statistical estimation algorithms, the topological structure of the schemas is pre-designed manually (Roy, 2005b).

4.6 Integrating Semantics and Syntax

So far we have presented mostly models of language evolution, i.e. how a language can emerge from scratch, or models of agents interacting with their environment and forming meanings to which a later language can ground. This could be glossed a "bottom-up approach".[18] Important insights for meaning formation can also be gained the other way round (or top-down) from natural language processing (NLP) systems. These systems usually require several levels of analysis, from phonological, morphological, syntactic, semantic, contextual to pragmatic, each with its own representational structures. Some of these analyses work sequentially, others could co-operate in reducing indeterminacies in parallel.

[18] With the exception of ELIZA and SHRDLU, see Section 4.1.1.

4.6.1 SAPFO: Frames and Semantic Networks

The system SAPFO written by Páleš (1994) is a noteworthy example of NLP system for Slovak language. The system gets a textual input – a sentence, it then translates the sentence into deep semantic structures and back to produce paraphrases[19] of the original sentence. This task is motivated by a commonly used criterion of understanding in humans: for example, a student is asked to reproduce the learnt content in his/her own words to prove that he/she understands the matter.

The system SAPFO, written in PROLOG, consists of lexical *data* (morphologic, lexical-semantic, syntactic, synonymic, phraseologic etc. dictionaries in the form of tagged words, tables, predicates and frames), and *algorithms* and general rules. The data incarnate an unusually detailed linguistic analysis of Slovak language. Although the program works on the level of formal manipulations, it produces semantically interesting behavior achieved by the co-operation of implemented analyses. The resulting deep semantic structures take the form of case frames (Minsky, 1975) or semantic networks (Russell and Norvig, 1995; Návrat et al., 2006).[20] The system can construct very complex semantic representations, e.g. a semantic network that captures the meaning of the following paragraph:

> "*John took Jenny to the silver lake. He wanted to show her the dignity of swans bathing in sun rays on the water surface. But the swans have flown. Sun shined and the water in the lake withered.*" (Páleš, 1994, p. 220).

Evaluation. The system SAPFO incarnates substantial amount of knowledge of relations between various semantic and syntactic elements. If interpreted by a human, the sentences produced by SAPFO have similar meaning as an original to-be-paraphrased sentence. However, understanding of this meaning is not intrinsic to the system, which only manipulates uninterpreted symbols interwoven in networks of mutual relations. We could say that it uses a dictionary-like structuralist approach to meaning conceived as the place of a word in relation to other words (de Saussure, 1916/1974). There is no coupling with the external world, and the whole knowledge is pre-programmed by the human designer.

In spite of that, SAPFO constitutes an important complement to the previously described models of grounded meaning creation. While SAPFO

[19]By paraphrases we mean sentences that can have different surface structure but the same meaning as the original sentence.

[20]These two forms are equivalent: semantic networks that allow for better orientation and visualization have one-to-one correspondence to frame-like structures of attribute-value pairs (Russell and Norvig, 1995, p. 298).

lacks grounding of elementary tokens, the other models lack compositionality and constraints on how to combine elementary meanings. In a sense, SAPFO shows a direction how to enhance the grounded systems with more complex meanings. The other way round, experiences with embodied systems suggest that (and how) the elementary tokens should be grounded in some embodied sensorimotor structures. We will describe a system going in this direction in the next section.

4.6.2 ECG: Schemas and Constructions

Full-fledged language understanding goes beyond meanings of single words: semantic relations between meanings of parts of an utterance (*who did what to whom*) are encoded by grammatical means such as word order or case markers.

Relations between syntactic elements and their effect on semantic interpretation can be expressed within *Embodied Construction Grammar* – ECG framework (Bergen and Chang, 2003). In ECG, linguistic knowledge repository consists of frame-like structures called *constructions* that link form elements (constraints on sound, surface word form, syntax, etc.) with meaning elements represented by *embodied schemas* — cognitive structures generalized over recurrent perceptual and motor experiences. The simplest embodied schemas can be conceived as a list of roles that allow external structures (including other schemas as well as constructions) to refer to the schema's key variable features, providing a convenient degree of abstraction for stating diverse linguistic generalizations (see Figure 4.8). More importantly, schema roles serve as parameters to more detailed underlying structures that can drive active simulations.

Constructions support a language understanding process modeled as having two distinct phases:

analysis – utterances are first analyzed to determine which constructions are involved and how their corresponding meanings are related: this process mostly utilizes constraint satisfaction techniques and its result is a *semantic specification* – the network of interconnected constructions, see Figure 4.9.

simulation – embodied schemas of the semantic specification are then simulated to produce inferences. Simulation itself relies on an active structure called an executing schema (or *x-schema*) that captures hierarchical structure, sequential flow, concurrency and other properties of motor control and event structure in general. Results of simulation are used to update a belief network representing the current context.

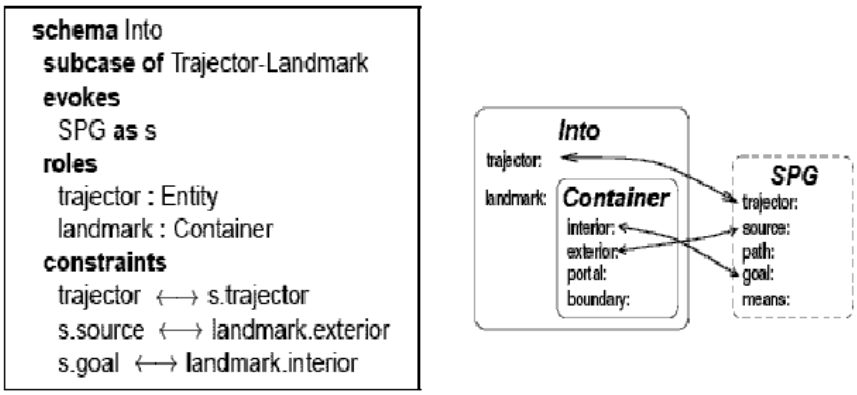

Figure 4.8: The Into schema, defined using the ECG formalism (left) and informally depicted as a set of linked schemas (right). Into is defined as a subcase of Trajector-Landmark that evokes an instance of the SPG schema (shown with a dashed boundary at right). Type constraints on roles require their fillers to be instances of the specified schemas, and identification bindings (⟷) indicate which roles have common fillers. The picture is taken from Bergen and Chang (2003).

Evaluation. ECG framework is primarily designed for language understanding, such as the metaphorical analysis of newspaper articles (Narayanan, 1997), but it was also used in the model of the acquisition of early phrasal and clausal constructions by Chang (2004). As the framework is rather complex, it has not yet been implemented as a whole. However the results of its partial implementations are promising and show the direction in moving from meanings of single words to the sentence-level embodied semantics.

4.7 Neural Networks

With the recent onset of non-invasive brain-imaging methods (e.g. PET, fMRI), neuroscience started to play quite a prominent role in nowadays cognitive science. It has not only brought valuable insights into brain functioning, but has also put new questions that should not be ignored by any theory or model of human cognition: is the proposed model/theory neurally plausible? Is it in accordance with known relevant facts about human brain? What neural mechanisms and representations could be behind the studied phenomenon?

Many connectionist models have addressed these questions in the language domain; for overview, see e.g. Farkaš (2005). In the following section,

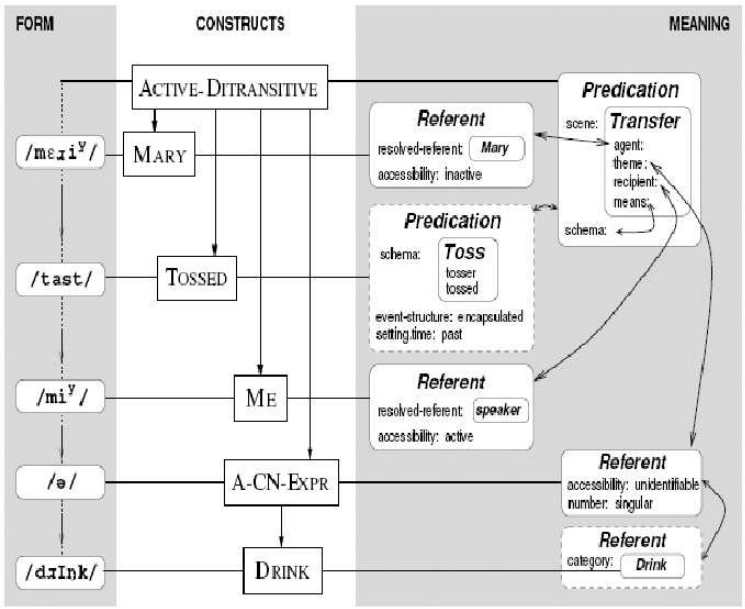

Figure 4.9: A depiction of a constructional analysis of the sentence "Mary tossed me a drink." in ECG framework. Constructs involved are shown in the center, linking elements and constraints in the domains of form and meaning; schemas are shown as rounded rectangles. The picture is taken from Bergen and Chang (2003).

we introduce an example of a connectionist model DevLex focusing on the nature of internal representation of lexical meanings and its development. A comprehensive connectionist approach to modeling human semantic cognition can be found in Rogers and McClelland (2004).

4.7.1 Lexical Development

The connectionist model of early lexical development DevLex (Li et al., 2004) tries to overcome limits of many current neural network models of language acquisition in the following aspects: First, it uses real corpus-based speech data that correspond to actual language use and not some artificially generated or small-set vocabulary. Second, it uses self-organization and does not need a supervised training (e.g. back-propagation). Third, it can cope with continuously growing lexicon and is scalable.

The network consists of two *growing self-organizing maps* (Farkaš, 2003) – a semantic S-GMAP and a phonological P-GMAP, that are connected via associative links trained by Hebbian learning. DevLex operation involves three processes: (1) formation of distributed word representations (both phonological and semantic), (2) GMAP organization, and (3) formation of associative links between form and meaning. The second and third processes occur simultaneously. The first process can be thought of as the process in which the child extracts phonological and semantic information from lexical contexts (sentences) during listening. It is done independently and its resulting form and meaning representations serve as input to the second and third processes. In this text, we will only focus on the semantic part, see Figure 4.10.

Semantic representations are built up gradually during the development of the lexicon in two qualitatively different ways. The first set of representations is constructed from word co-occurrence probabilities in the input corpus using the word co-occurrence detector (WCD) recurrent network (Farkaš and Li, 2001). The second set is static[21] and consists of binary vectors of semantic features of each word derived from the WordNet database by special feature-extracting routines (Harm, 2002).

The experiments with DevLex focused on development of internal representations of lexical meanings in GMAPs and on reproducing several phenomena observed in early language acquisition by children, e.g. lexical confusion and age-of-acquisition effects. 500 word input lexicon was divided into sets corresponding to major developmental stages of child language acquisition. While feeding the network with these input sets, internal organization of lexical representations in GMAPs was inspected. The results revealed that the

[21] In the sense that it does not evolve with the growing vocabulary.

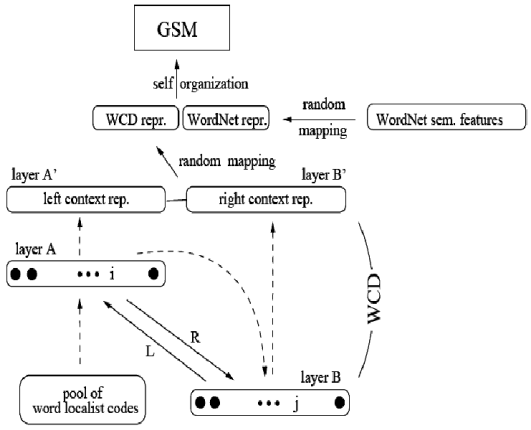

Figure 4.10: The semantic part of DevLex network. The bottom half represents the word co-occurrence detector (WCD), the upper part represents the random mapping followed by self-organization of the growing semantic map (GSM). The picture is taken from Farkaš (2003).

internal representation captures both syntactic and semantic relations between words. Nodes that represented words belonging to the same syntactic category (nouns, verbs, etc.) or having related meanings (e.g. *now*, *then* or *up*, *down*) tended to be close to each other and/or form clusters. Moreover, the representation changed in time, providing important insights into the nature of child lexical development (for details, refer to the original source).

Evaluation. First of all, the presented model addresses the important issue of neural plausibility of modeling the organization of semantic representations from the developmental point of view. Also, it validates the emergentist alternative to the nativist assumption that lexical categories are hardwired in the brain. Learning in the system is in the large part based on structuralist relations between words co-occurring in the corpus.[22] The lack of grounding in the real world is partially remedied by supplying the semantic features of words from an external static source. Self-organization could facilitate grounding, because it enables the autonomous development of natural categories. However, this could only happen if the system somehow interacted with the real world in a feedback loop, e.g. by producing some behavior that would in turn influence its own inputs. In the next section, we

[22]Neural networks can be surprisingly good in inducing grammatical information implicit in transitional probabilities in the corpus. Recurrent neural networks show the *architectural bias* to meaningful next-symbol predictions even before training. Training causes reorganization of the network's state space according to grammatical categories (Čerňanský et al., 2007).

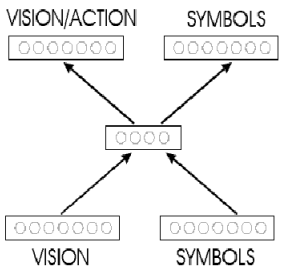

Figure 4.11: A typical dual-route architecture for connectionist models of symbol grounding. The picture is taken from Cangelosi (2005).

present several connectionist models that ground language in representations formed by acting in the world.

4.7.2 Integrating Perception, Action and Language

The importance of links between perceptual, sensorimotor and cognitive abilities for symbol grounding has been supported both theoretically and experimentally (Pecher and Zwaan, 2005). Connectionist models of symbol grounding often employ dual-route architecture (see Figure 4.11) that typically involves both visual input (e.g. retina projection or visual feature list) and linguistic input (e.g. localist or graphemic/phonetic encoding of symbols). The output layer has symbolic units for representing words (e.g. with a phonetic encoding of the lexical items), and either a categorical representation of input stimuli (e.g. a localist node for each category, or a visual representation of category prototypes) or representation of a desired action (e.g. values of joint angles of an arm). All input and output layers are connected via a shared hidden layer. The route from visual input to symbolic output is used for language production tasks, such as naming of the object represented in the visual scene or its category. The route from linguistic input to visual/categorical/motor output is used for language understanding tasks. The two other possible routes are used for categorization and sensory-based action (the route from visual input to categorical/motor units) and for linguistic imitation (from linguistic input to symbolic output) (Cangelosi, 2005).

In the model of Cangelosi (1999b), genetic algorithms and neural networks are combined together for studying the emergence of compositional lexicon in an ecological setting: a population of 80 organisms lives in a virtual environment, where each organism performs a foraging task by collecting "edible mushrooms" and avoiding "toadstools". There are 6 types of mushrooms in

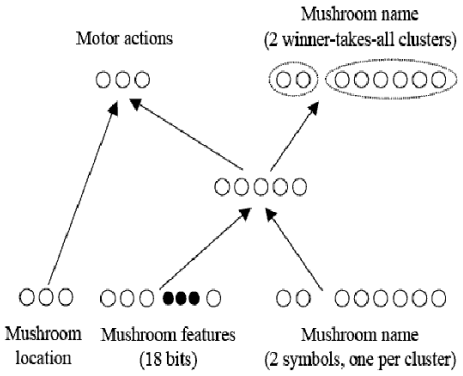

Figure 4.12: The neural network that controls the behavior of the foraging organism. In the input layer, 3 units encode the location of the closest mushroom and 18 units encode their binary features. Eight input units are used for the 8 symbols (words) used for naming mushrooms. The network has 5 hidden units. In the output layer, 3 units control the organism's behavior (movement and identification of mushroom category), and 8 units are used to encode the mushroom names. These symbolic output units are organized in two clusters of competitive winner-takes-all units (one cluster of 2 units, the other of 6 units). Since only one unit per cluster can be active, each mushroom will be named using two symbols. The picture is taken from Cangelosi (1999a).

the environment: three edible and three poisonous. To gain fitness, organisms have to avoid toadstools and identify the type of edible mushrooms and eat them.

Behavior of each organism is driven by a dual-route network (see Figure 4.12) that enables performing actions (based on mushroom location and perceptual features and/or linguistic input from other agents) and naming (either as imitation of a linguistic production of another agent or naming the mushroom for another agent based on the mushroom's location and perceptual features). At the end of their lifetime, the fittest 20 organisms are selected and reproduce 20 offspring each. The new 80 organisms live together with their 20 parents that serve as speakers and teachers for naming the mushroom categories and the actions to take. Although at the beginning of evolution the lexicon is totally random and meaningless, toward the end of the evolution the agents are able to evolve shared compositional languages. Inspection of representation on the shared hidden layer revealed that language helped categorization in increasing inter-cluster and decreasing intra-cluster distances (in comparison to the stage with purely sensori-

motor task without any language). The same result has been confirmed in the dual route architecture based experiments of Cangelosi and Parisi (2001) and Mirolli and Parisi (2005).[23]

Evaluation. The presented models involve both performing actions in a simulated environment and linguistic tasks. The language learning is connected to sensorimotor activities via a shared hidden layer in the dual route architecture. Internal representations of nouns (object names) covary more with perceptual features, while internal representations of verbs (action names) covary more with motor control parameters (Cangelosi and Parisi, 2004). Meanings, implicitly represented in connection weights of the networks are private and created individually in each organism. Unlike in the related model of Kvasnička and Pospíchal (1999) described in Section 4.3.2, the fitness of organisms in the presented model of Cangelosi depends on successful performance of a task and not on the comparison of private meanings. Although the presented models are deliberately simple, they illustrate a correct connectionist approach to symbol grounding.

4.8 Corpus-Based Meanings

In Section 4.7.1, we presented the connectionist model DevLex with semantic representations partially based on transition probabilities of lexical units in the training set corpus. Although not sufficient for language grounding in the real world (see the Chinese Room metaphor in Section 1.1.1), distributional and contextual information is an important cue to word meanings (Li et al., 2004). This is impressively demonstrated by chatbots based on contextual corpus search that we shall describe in the following section.

4.8.1 Jabberwacky chatbots

A chatbot is a computer program designed to simulate an intelligent conversation with human users. The classical example of a chatbot is ELIZA (see Section 4.1.1). Many chatbots are based on recognizing keywords in the human user's input and answering according to pre-programmed rules. Jabberwacky[24] written by Rollo Carpenter is an example of a chatbot based on different principles; there are no fixed rules programmed into the sys-

[23] Unlike in the former experiments, the latter one did not involve genetic evolution. The one-generation neural network was trained by backpropagation learning.

[24] Available online at `http://www.jabberwacky.com`

tem[25] and it operates entirely through user interaction. The system maintains a large database of all previous conversations and attempts to use this information to find the most appropriate response in the current context. The search is data-centric, probabilistic and statistical, yet at the same time chaotic, which means that tiny differences in context can lead to huge differences in answers. It is not based on any single recognized AI technique such as Markov chains or neural networks, but it is a complex layered set of heuristics that produce results through analyses of conversational context and positive feedback (Icogno, 2007).

The success in giving an impression of real thinking depends on a large-enough database. Having been online on the world wide web since 1997, Jabberwacky has recorded more than 13 million conversations. Jabberwacky reflects back what it had learned from its conversation partners. In this way, it can use jokes, idioms, word games, slang, and even speak foreign languages.

Two recent instantiations of Jabberwacky – George and Joan have won the Loebner prize in 2005 and 2006. The Loebner Prize is an annual competition that awards prizes to the chatbot considered by the judges to be the most humanlike of those entered. The format of the competition is that of a standard Turing test (see Section 1.1.1). Unlike previous version based on purely textual form of communication, George has a 3D visual "avatar" appearance with a variety of facial expressions and the ability to understand and respond to others using human speech. The commercial interest in George suggests the future areas of application of such chatbots/avatars in education, interactive entertainment systems, advertising and sales.

Evaluation. Although Jabberwacky's conversations can be interpreted by humans as having some semantic content, i.e. as being "about something", this content is extrinsic and Jabberwacky knows nothing about it. Hence, its linguistic knowledge is not grounded in the real world (see Section 1.1.5). However, from a different point of view, Jabberwacky is situated in the world of conversational sequences, where it learns from scratch to react appropriately in various contexts. The knowledge of what is "appropriate" is encoded in the recorded history of reactions of the chatbot's human partners. In line with Brooks (see Section 1.1.4), we can view Jabberwacky's intelligent behavior as an emergent effect of its interactions with the environment, where the knowledge is distributed both in its architecture and the environment.

[25] That is, no rules of grammar or conversation conduct; rules for learning and contextual database retrieval are naturally build in the system.

4.9 Summary

In this chapter, we reviewed main approaches to representation of meanings in artificial systems (computational models, programs, agents and robots). We evaluated the models with respect to issues of symbol grounding, learning and interacting with the environment. The evaluation results are summarized in Table 4.1.

Evaluated system	Representation of meanings	Environment: Real/Simulated	Meanings: Given/Learned	Scope: Objects/Verbs	Supports syntax?	Allows telepathy?
ELIZA (Weizenbaum, 1966)	rules	–	G	–	Y	N
SHRDLU (Winograd, 1971)	procedures	S	G,L	O,V	Y	N
Shakey (Nilsson, 1984)	predicates	R	G,L	O,V	Y	N
ILM (Kirby and Hurford, 2001)	predicates	–	G	O,V	Y	Y
(Oliphant, 1997)	scalars	–	G	–	N	Y
(Kvasnička and Pospíchal, 1999)	vectors	–	G	–	Y	Y
Talking Heads (Steels, 1999)	discrim. trees	R	L	O	N	N
(de Jong, 2000)	situation concepts	S	L	O,V	N	N
(Vogt, 2002)	prototypes	R	L	O	N	N
Neo (Cohen et al., 1996)	redescriptions	S	L	O,V	–	–
(Cohen, 1998)	dynamic maps	–	L	V	N	–
Ripley (Roy et al., 2004)	semiotic schemas	R	G,L	O,V	Y	N
SAPFO (Páleš, 1994)	semantic networks	–	G	O,V	Y	N
ECG (Bergen and Chang, 2003)	constructions	S	G,L	O,V	Y	N
DevLex (Li et al., 2004)	NN configuration	–	L	O,V	N	Y
mushroom foragers (Cangelosi, 1999b)	NN configuration	S	L	O,V	Y	N
Jabberwacky (Carpenter, 2007)	corpus relations	S	L	–	Y	N

Table 4.1: Summary of evaluation of meanings in the reviewed artificial systems. By 'telepathy' we mean direct access to internal semantic representation of a communication partner. Y = yes, N = no, '–' = irrelevant, other acronyms correspond to capital letters in column headers.

Chapter 5

Understanding Revisited

In previous chapters, we went through formal theories of meaning, an evolutionary view on onset of understanding in living organisms, and main types of semantic representations used in computational models and other artificial systems. In this chapter, we summarize our own notion of meaning and understanding.[1] To avoid non-productive debates about whether the machine understanding is a "true" one similar to that of human beings, we try to use neutral descriptions that would not exclude machines by definition. This is in line with similar efforts to define life (Csontó, 2001; Csontó and Palko, 2002) and consciousness (Wiedermann, 2006, 2007) in such a general way that human life and human consciousness are just their possible instantiations. Extended elaboration of ideas presented in this chapter can be found in Šefránek et al. (2007).

Language competence is not an isolated module, but a result of many cooperating cognitive processes. Understanding does not begin with language – in fact, lexical meanings are part of the conceptual system that has been shaped by experiences with the surrounding world. We can also talk about understanding on preverbal level, in the sense of "understanding the world and its laws". Hence, we can draw lessons from studying preverbal stages of ontogeny and phylogeny: studying sensorimotor intelligence of infants and animals.

Understanding is inherently individual: each organism has its own Umwelt determined by its purpose in its environment, its embodiment and perceptual/motor abilities and its interactional history. What is imperceptible or just a meaningless noise for one organism, can be interpreted (understood) by another organism as meaningful. A sign is understood as having a seman-

[1] Because most of the statements in this chapter had already appeared in the previous chapters, we present them without references if they had been supported by references elsewhere.

tic content for an observer, if the observer behaves toward it in accordance with this content (van Gulick, 1988). This view is interactionist. Meanings cannot be transferred directly from one organism to another; they can only be inferred from behavioral interactions.

Hence, we can only talk about understanding in the context of the environment that the organism is situated in. The organism is coupled with its environment by perception and action. This coupling creates a feedback loop that is a necessary condition for any adaptive behavior, including learning.

Elementary forms of understanding can be found in organisms that (at least implicitly) categorize the world by producing different behavioral responses for different classes of their perceptual inputs. We say that these organisms possess cued representations of the categories. Sophisticated understanding required for the language use is based on detached representations that can be retained, retrieved and processed independently of external triggers from the current environmental context of the organism.

These principles can be applied to study meanings in artificial systems. Again, an artificial system must be situated in an environment and interact with it by sensorimotor activities (perception and action). The environment does not have to be physical: autonomous software entities (agents) can "live", i.e. sense and perform actions, in a virtual environment, e.g. search in databases or negotiate e-commerce transactions.

Non-trivial environments are dynamic and changing in time. This has important consequences for the design of "understanding" agents. First, besides static objects, the agents must be capable to represent dynamic characteristics of the world, such as changes, actions, their consequences and events. Second, because the environment is open, all possible meanings cannot be anticipated and the agents should learn. Learning (construction of meanings) should be incremental and continuous.

Basic distinction can be made between categorical and propositional meanings. Categorical meanings are more basic and consist in factorization of continuous input space into discrete number of classes. Members of each class/category are treated equal with respect to some purpose. We say that categories are *environmentally valid* (or natural), if they reflect distributional characteristics of properties of the environment. Environmentally valid categories form clusters with high inter-cluster and low intra-cluster differences. Such categorization can be constructed by unsupervised learning methods.

Categories are *ecologically valid*, if they reflect some pragmatic effect on the agent, e.g. division of mushrooms to edible and poisonous regardless of their perceptual similarity. Such categorization can be arrived at by utilizing feedback from the environment (in this sense, learning is supervised).

Many meanings can be innate, if they had been vital for survival on the

evolutionary timescale. Other meanings are constructed by observing the environment and consequences of one's own actions. Yet another meanings are transmitted culturally; this is when language enters the scene. Empirical findings from child language acquisition suggest that language has an important influence on meaning formation process. The language influence is necessary for meaning coordination or social symbol grounding: as meanings are constructed individually and cannot be transferred directly, they must be attuned to each other via linguistic means.

Propositional meanings take the form of assertions that can be true or false and capture the agent's beliefs and/or causal and relational knowledge about the properties of the environment (including the agent itself). Agent can use them for reasoning, predictions, planning and overall orientation in the world. In connection to language, they correspond to sentence-level semantics and are often centered around verbs.

This chapter concludes the first, theoretical, part of this thesis. Here we formulated our notion of understanding and meanings with the emphasis on grounding of lexical meanings in sensorimotor and social (linguistic) interactions. The second, computational, part of the thesis is dedicated to our own original proposal of semantic representation amenable to autonomous construction/acquisition, and to experiments aimed to validate the proposed representation.

Computational Models

Chapter 6

Methodology

We have decided to study the problematics of meanings and their origin by means of synthetic modeling methodology. That is, a theory we will formulate must be validated by experimenting with its computational implementation. The theory should stem from and be consistent with our general notion of meaning and understanding, as formulated in Chapter 5. Here we briefly state the methodological commitments implied by this notion.

6.1 Commitments

1. Our notion of meanings applies not only to linguistic humans, but also to preverbal living organisms and artificial systems.

2. Meanings should not be given before-hand. Therefore, the proposed theory should explain the mechanisms of meaning emergence, both on the individual and social levels.

3. Meaning construction mechanisms should be based on interactions with the environment and other agents.

4. In linguistic interactions, telepathy (direct access to internal representation of the other communication partner) is not allowed. In bootstrapping language from scratch, agents can use externally observable behavioral hints (such as pointing or gaze following) to narrow down the communication context.

5. Importantly, possible meanings should not be limited to object categories; they should also include properties, relations, dynamic changes, situations and events.

6. All the aspects of the meaning representation and meaning construction must be formulated rigorously enough to allow for their computational implementation and experimental validation.

6.2 Experimental Plan

Adhering to these commitments, we take the following steps:

1. We propose a similarity-based semantic representation of various types of meanings.

2. We propose individual and social mechanisms of autonomous construction of such semantic representations.

3. We implement the hypothesized mechanisms in computational models and analyze the results of simulations.

4. Also, we study the dynamics of meanings in a computational model that includes inter-generational transmission of meanings by iterated learning.

Chapter 7

General Framework of the Models

We have designed, implemented, simulated and evaluated several novel and original computational models of different mechanisms of meaning construction. In this chapter, we start with description of the general framework of these models.[1]

A model typically consists of one or more agents and a simulated environment. The agents are coupled with the environment by processes of sensing and acting. In each time step, each agent senses its environment, updates its internal representation and can communicate or perform other actions, depending on a particular application (Takáč, 2005b).

Internal processes of agents operate on different kinds of representation, which can be described on the following levels (see Figure 7.1):

Perceptual level. This level is an interface between the external environment of the agent and higher levels. It is the product of the agent's perception/sensation process. This level of representation is iconic in the sense of Harnad (1990).[2] In embodied agents, it could represent the signal from the agent's sensors pre-processed by low-level perceptual routines. In software agents, it represents the input data the agent operates with, translated to the description processable by the conceptual level.

Conceptual level. This is a level of categories/concepts. Each concept is represented by an *identification criterion* – the function that maps a

[1] Not all the features described here are applicable in all the models. Individual deviations from this general framework will be emphasized in the chapters describing particular models and experiments.

[2] See Section 1.1.5.

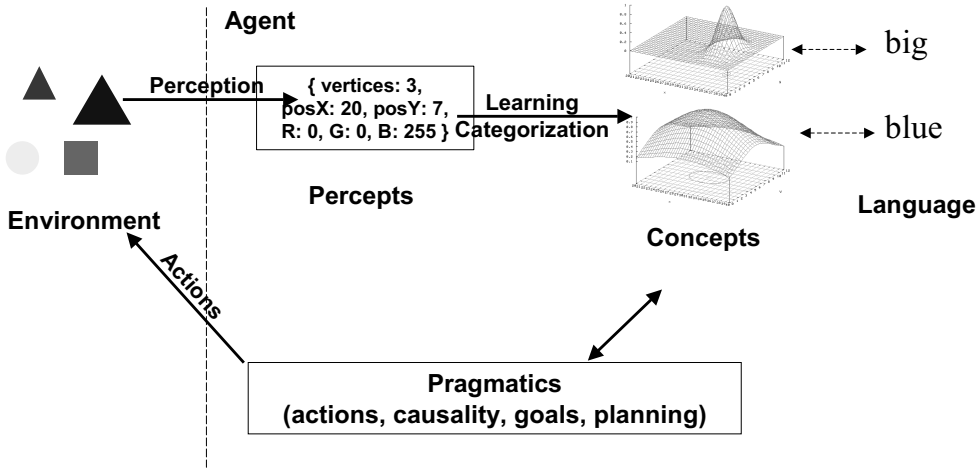

Figure 7.1: The cognitive architecture of the agent includes perception, representation, language and pragmatic modules.

perceptual[3] input to a numeric activity value expressing to what extent the input is an instance of the concept.

Language level. The agent's identification criteria are private and are not directly transferable to other agents. The agents communicate by exchanging conventionally established signals of the language level. The meanings of the signals are the perceptually grounded criteria of the conceptual level. The communication can be successful, only if the private meanings of the agents are sufficiently similar. This occurs, if the agents use similar concept formation mechanisms and have similar experiences in the shared environment.

Pragmatic level. On this level, the agent plans and achieves its goals in the environment. It uses representations of causal knowledge about its actions and their consequences in the form of cross-categorical associations of criteria, own goals as desired situations, and plans as sequences of actions leading from the current situation to a desired one.

7.1 Environment

The agents are situated in a simulated environment together with other entities (objects). The environment is dynamic and open, in the sense that the

[3]More complex criteria of situations and events operate on outputs of other criteria (see Sect. 8.3.3).

objects (including the agents themselves) can appear or disappear or change their properties in discrete time steps.

Technically, all objects are described by frames of attribute–value pairs (see the next section) and the values of attributes can change in each time step. In case the agents perform actions, the consequences of the actions on the objects are evaluated and realized by the environment module, which resolves potential conflicts between agents and applies effects of (simple simulated) physical laws, e.g. if an agent tries to lift an object that is too heavy, the environment module ignores this action and the properties (altitude, in this case) of the object, will not change.

7.2 Perception

In each time step, each agent perceives objects in its immediate environment.

Let $\left\{o_1^{(t)}, \ldots, o_m^{(t)}\right\}$ be objects[4] in the vicinity of an agent A at time t. Then the perceptual input (perceived *scene*) $S_A(t)$ of an agent A at time t is the set of frames

$$S_A(t) = Sel_A\left(\left\{o_1^{(t)}, \ldots, o_m^{(t)}\right\}\right) = \left\{f_{i_1}^{(t)}, \ldots, f_{i_n}^{(t)}\right\} \ ,$$

where

$$\{i_1, \ldots, i_n\} \subseteq \{1, \ldots, m\}$$

and each $f_{i_k}^{(t)}$ is a perceptual image (frame) of the object $o_{i_k}^{(t)}$.

We can see that the selection function Sel_A modeling the individual perceptual abilities of the agent A first selects objects $o_{i_1}^{(t)}, \ldots, o_{i_n}^{(t)}$ visible to the agent and then projects each of them to a corresponding perceptual frame.

Formally, a perceptual frame f is characterized by the set of attributes A_f and the real-valued attribute accessor function $h_f : A_f \to \mathrm{R}$. (In the text, we will use a more conventional notation $f.a$ instead of $h_f(a)$).

In general, the sets of attributes (and their values) of an object $o_{i_k}^{(t)}$ and its perceptual image $f_{i_k}^{(t)}$ do not have to be identical. Some attributes may not be perceivable by the agent,[5] others may be transformed by pre-processing routines.

The scene composed of perceptual frames represents all the input relevant to the agent, e.g. objects in the physical environment of the agent, incoming data, or the agent's "proprioceptive" input (values of internal variables,

[4] All objects (including the perceiving agent itself) are described by frames.

[5] For example, if the perceived object is another agent, its internal properties, such as its hunger level, may be visible just to that agent, but not to other agents.

parameters of operations performed, position of an arm, etc.). The names of attributes have no special meaning for the agent, except for establishing a correspondence of attributes with the same name in different frames. The values of attributes are real numbers.

Frames have a long tradition in artificial intelligence (Minsky, 1975) and cognitive semantics (Fillmore, 1982); they are easily readable by human observer, general enough to describe various data structures, and they can be implemented in the spirit of structured connectionism (Shastri et al., 1999; Feldman, 2006).

Using frames or other arbitrary amodal symbols as semantic representations has been criticized by Barsalou (1999). Barsalou defines representation as amodal, if its internal structure bears no correspondence to the perceptual states that produced it. However, frames in our architecture are general enough to provide means for modal (iconic) as well as amodal projections of the external objects. They model the structures resulting from a low-level preprocessing of the perceptual input. For example, a frame can represent an array of intensity values of retina or a camera image, or a scene segmentation to perceptual characteristics of particular objects.

In the models presented in this thesis, the problem of perceptual preprocessing is abstracted away.[6] Hence, all agents receive the same perceptual input – directly the frames of objects $o_1^{(t)}, \ldots, o_m^{(t)}$ (i.e., the selection function Sel_A of each agent is the identity function). However, each agent can categorize and represent the same perceptual input in its own individual way.

7.2.1 Perception of Changes

The ability to perceive and represent changes is very important for any agent operating in a dynamic environment. In a continuous world, even 4-month-old children can eye-track moving objects and develop the concept of *object continuity* (Johnson et al., 2003), which is a necessary condition for noticing changes of objects.

As the time is discrete in our model, we must ensure that the agent does not perceive scenes in subsequent time steps as independent, but as the sets with established correspondences between frames representing percepts of the same object at different times (to be able to track individuals in time).

We manage this be refining the definition of the *scene* from the previous

[6] Autonomous construction of discrete perceptual structures from raw continuous sensory readings was modeled e.g. by Rosenstein and Cohen (1998) and Kuipers et al. (2006).

section to a set

$$S_A(t) = \left\{ \left(f_1^{(t)}, f_1^{(t-1)} \right), \left(f_2^{(t)}, f_2^{(t-1)} \right), \ldots, \left(f_n^{(t)}, f_n^{(t-1)} \right) \right\}$$

of percepts linked with their one-step history. If an object just appeared on the scene at time t, its history frame $f^{(t-1)}$ will be assigned a special value $\perp$. If an object had been present on the scene in the time $t-1$ and now disappeared, $f^{(t)} = \perp$.[7] Otherwise both $f^{(t-1)}$ and $f^{(t)}$ are standard frames as defined in the previous section (the scene does not contain pairs with $\perp$ on both positions).

7.3 Conceptual Representation

The proposal of representation of various kinds of concepts is the crucial part and main contribution of this thesis. Therefore we describe it in detail in a separate chapter (Chapter 8).

7.4 Language

The agent's lexical knowledge is stored in the form of bi-directional associations of meanings with words. The words are arbitrary strings of characters; the meanings are identification criteria of the conceptual level.

7.4.1 General Case

In general, one meaning can be associated with several words (synonymy) and vice versa (homonymy). Each association has a strength (expressed by a positive real number).

Understanding a Word

Out of the meanings associated with a word in the lexicon, the agent selects a meaning with the highest *confidence*. Confidence is a strength of the association divided by the sum of strengths of all associations of this word with meanings (Smith, 2003b).

[7] Identification criteria *appeared* and *disappeared* can be based on detecting $\perp$ on the respective position in the input pair.

Verbalizing a Meaning

To verbalize a meaning, the *introspective obverter* strategy (Smith, 2003b) is used: out of the words associated with the meaning, the agent selects the one that it itself would best understand as the meaning. If it does not have such a word, it can invent a random one.

Learning

Learning is managed by manipulating the strengths of word-meaning associations based on the feedback about success in communication (Steels, 2000) or on cross-situational co-occurrences of words with meanings (Smith, 2005a).

7.4.2 One-to-One Associations

The language learning assumptions of our model (see Section 10.1.2) cause that all associations will be one-to-one. This simplification makes learning easier and allows us to focus on more interesting issues of relations between meanings (represented by identification criteria) and the environment. Relating the language and meanings to the world is in the domain of pragmatics.

7.5 Pragmatics

7.5.1 Relation to the World

A practical *use* of the learned concepts and lexical expressions depends on particular applications. Here we define several functions formalizing the relations between concepts, words and the external world (see Figure 7.2).

We start with the notion of *focus*. The focus is used for determination of a part (or aspect) of the scene that is denoted by a linguistic expression or a particular concept.

Formally, the focus ϕ is a projection of the scene

$$S^{(t)} = \left\{ \left(f_1^{(t)}, f_1^{(t-1)} \right), \left(f_2^{(t)}, f_2^{(t-1)} \right), \ldots, \left(f_n^{(t)}, f_n^{(t-1)} \right) \right\}$$

that selects a particular (ordered) k-tuple of perceptual frames

$$\phi\left(S^{(t)}\right) = \left\langle f_{i_1}^{(t)}, \ldots, f_{i_k}^{(t)} \right\rangle \tag{7.1}$$

or a particular (ordered) k-tuple of pairs of frames

$$\phi\left(S^{(t)}\right) = \left\langle \left(f_{i_1}^{(t)}, f_{i_1}^{(t-1)} \right), \ldots, \left(f_{i_k}^{(t)}, f_{i_k}^{(t-1)} \right) \right\rangle \ . \tag{7.2}$$

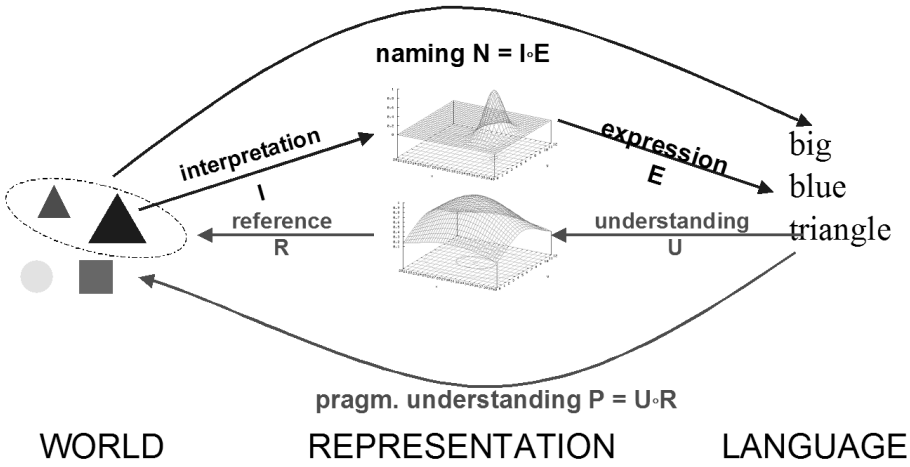

Figure 7.2: Pragmatic functions relating concepts and language to the world.

We shall call the result of the projection function a *referent* of the focus.

The focus restricts the whole scene to a part that forms an input to a particular *identification criterion*. Identification criteria are functions that represent concepts. They take various types of input arguments and return a value that determines the degree of the input's membership in the category (see Section 8.3).

Criteria of objects and/or properties[8] take as an input argument a perceptual frame of one object, i.e. a focus projection of the scene with $k = 1$ in Eq. (7.1). Binary relational criteria operate on pairs of perceptual frames ($k = 2$ in (7.1)) and criteria of situations operate on subsets of the scene (a general case of (7.1)). Change criteria operate on a pair of frames of the same object in different times ($k = 1$ in (7.2)) and event criteria express multiple co-occurring changes (a general case of (7.2)). Also, multiple criteria can be aggregated to hierarchies where higher-order criteria operate on output activities of other criteria.

Let us assume that the agent has acquired a lexicon of one-to-one associations $\mathcal{L} \subset W \times C$, where W is a set of learned words and C is a set of concepts (criteria). Let S be a scene and ϕ be a focus. Then we define:

Understanding. The function $U : W \rightarrow C$ returns a criterion r that is the meaning of the word w. For $w \in W$,

$$U(w) = r, \text{ such that } (w, r) \in \mathcal{L} \ .$$

[8]I.e. monadic criteria, see Section 2.1.4.

Expression. The function $E : C \to W$ expresses the criterion r by the word w. For $r \in C$,

$$E(r) = w, \text{ such that } (w, r) \in \mathcal{L} \ .$$

Interpretation. The function I returns a set of concepts that a referent is an instance of, with the degree of membership determined by the threshold parameter θ. For a referent $\phi(S)$,

$$I_\theta(\phi(S)) = \{r \in C \mid r(\phi(S)) > \theta\} \ .$$

Naming. The function N is a composition $I \circ E$ and returns names of all categories that the referent is an instance of. For a referent $\phi(S)$,

$$N_\theta(\phi(S)) = \{w = E(r) \mid r \in I_\theta(\phi(S))\} \ .$$

Scene Interpretation. The function J returns a set of concepts that have at least one referent on the scene at the membership threshold level θ, together with foci determining their referents. For a scene S,

$$J_\theta(S) = \{(r, \phi) \mid r \in C,\ r(\phi(S)) > \theta\} \ .$$

Scene Description. The function D is a composition $J \circ E$ and returns names of all concepts found by scene interpretation at the membership threshold level θ, together with foci determining their referents. For a scene S,

$$D_\theta(S) = \{(w, \phi) \mid w = E(r),\ (r, \phi) \in J_\theta(S)\} \ .$$

Reference. The function R returns a set of foci determining the referents of the meaning r present on the scene S. Strictness of the membership is given by the threshold parameter θ. For a meaning $r \in C$ and a scene S,

$$R_\theta(r, S) = \{\phi \mid r(\phi(S)) > \theta\} \ .$$

Pragmatic Understanding. The function P is a composition $U \circ R$ and returns the set of foci determining the referents of the word w present on the scene S. For a word $w \in W$,

$$P_\theta(w, S) = R_\theta(U(w), S) \ .$$

The reference function can have a contrastive version $R^*(r, S)$ returning foci of referents on the scene S, for which r gives the maximum value (regardless of any threshold). Using this function in P enables the agent to understand also the contrastive use of words, e.g. if the agent heard a word "*big*" uttered along with the scene S containing only small objects, $P_{0.5}(\text{“}\textit{big}\text{”}, S)$ would return an empty set, while contrastive $P^* = U \circ R^*$ would return best matching objects, i.e. the biggest of the small ones.

7.5.2 Representation of Pragmatic Knowledge

One of the basic tenets of cognitive semantics (see Section 2.4.5) states that the very same representation that is used for language understanding is also used for reasoning and acting in the world. In Section 8.4.3 we will show how an agent can represent causal knowledge about consequences of its actions in the world. Once it has this kind of representation, it can plan sequences of actions presumably leading from its current situation to a desired one. A BDI[9] agent can be endowed with needs motivating its goals, which can be represented as identification criteria of desired situations. As the agent's knowledge is evolving and incomplete, the reasoning and planning is essentially non-monotonic and must include revisions.

[9]Belief - Desire - Intention (Bratman, 1987).

Chapter 8

Representation of Meanings

This chapter is dedicated to formal specification of representation of various kinds of categories and mechanisms of their construction (Takáč, 2006b,c, 2007b). We consider its content to be one of the main contributions of this thesis.

The goal is that artificial agents gradually learn to distinguish environmental properties and group entities similar in some respect into categories. Each category is represented by an *identification criterion*[1] – an activation function that returns, for some input, the degree of the input's membership in the category.[2] The possible inputs include a perceptual frame of one object (for criteria of objects and properties), perceptual frames of several objects (relational criteria), frames of the same object in different times (change criteria) and output activities of other criteria (compositional criteria of situations and events). The agents construct all their criteria from scratch by extracting common statistical properties of examples of categories encountered during their lifetime.

Extraction of statistical properties and categorization of novel examples is realized in *locally tuned detectors*, which form the core of every identification criterion. A locally tuned detector takes as input one frame and returns a value from the closed interval $[0, 1]$, expressing to what extent the frame is an instance of the category (1 means the best, prototypical example).

[1] In our previous works, we used the term *discrimination criterion*. In line with Harnad (1990), we have changed it, because the activation function actually returns the degree of identification of its input with the represented category (see the discussion in Section 4.4.5).

[2] Our notion of identification criteria, together with basic elements of the proposed semantics, is inspired by theoretical foundations laid by Šefránek (2002).

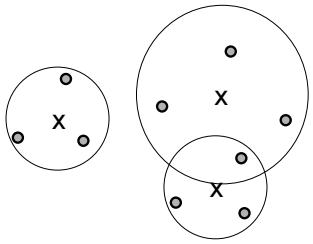

Figure 8.1: Categories represented by locally tuned detectors with thresholds do not have to be mutually exclusive and do not have to cover the whole input space. The points lying in the intersection of the two right circles belong to both categories, points outside all the circles do not belong to any of the represented categories.

8.1 Geometrical View on Categories

Locally tuned detectors have an intuitive geometric interpretation based on *conceptual spaces* (Gärdenfors, 2000), see Section 2.4.5. Perceptual frames defined in Section 7.2 can be viewed as vectors in the respective subspaces with dimensions determined by the attributes of the frames. A locally tuned detector should react with high activity to the convex hull of the vectors corresponding to examples of the represented category. The detectors represent categories with fuzzy boundaries (with their activity expressing the degree of category membership), but for practical purposes we can establish a decision threshold. In this case, the *receptive field* of a detector $r : D \to [0, 1]$ in the input space (domain) D, defined as the set $\Psi_\theta(r) = \{x \in D \mid r(x) > \theta\}$ for some decision threshold θ delineates a category. Locally tuned detectors have a high neural and biological plausibility (Martin, 1991; Balkenius, 1999; Hassoun, 1995).

The important difference between categories in a Voronoi-tessellated conceptual space (see Section 2.4.5) and categories represented by locally tuned detectors over a common space is that the latter categories do not have to be mutually exclusive and do not have to cover the whole input space (compare Figures 2.1 and 8.1).[3]

8.2 Construction of Locally Tuned Detectors

Functioning of the detectors is based on expressing common statistical properties of category examples in geometric terms and evaluating category mem-

[3]For comparison of conceptual spaces and receptive field based meanings, see also Takáč (2007d).

bership as a distance in a conceptual space.

Let us assume that the agent has to induce a detector from a sequence of example frames $\{f^{(1)}, \ldots, f^{(N)}\}$, each of which can be represented by a point

$$\left(f^{(i)}.a_1, \ldots, f^{(i)}.a_{|A_{f^{(i)}}|}\right)$$

in the respective conceptual space with dimensions corresponding to attributes $a_j \in A_{f^{(i)}}$. The induction is based on properties of values of attributes *common to all frames*. Hence, each frame $f^{(i)}$ is projected into a common subspace $\mathcal{A}$ with dimensions from intersection of all attribute sets $\bigcap_{i=1}^{N} A_{f^{(i)}}$. Attributes not present in every example are considered irrelevant for the category membership. From now on, we will represent the sample of the category as a set of projected vectors $\vec{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right)$ for $i = \overline{1, N}$ in the common space $\mathcal{A}$ of the dimensionality $n = |\bigcap_{i=1}^{N} A_{f^{(i)}}|$.

In line with Gärdenfors (2000), the geometric centroid computed as the mean vector of the sample set

$$\vec{p} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}^{(i)} \ .$$

will represent a prototype of the category.

The membership of a perceptual frame f in the category represented by a locally tuned detector $r_{\vec{p}}$ will be evaluated as an exponentially decaying function of the distance from the prototype (Shepard, 1987)

$$r_{\vec{p}}(\vec{x}) = \exp\left(-k \cdot d(\vec{p}, \vec{x})\right) \ , \tag{8.1}$$

where k is some positive constant, d is some metric and $\vec{x}$ is a projection of the frame f into $\mathcal{A}$ (if f cannot be projected because it lacks some attributes from $\mathcal{A}$, the detector returns 0).

The shape of the receptive field $\Psi_\theta(r) = \{\vec{x} \in \mathcal{A} \mid r(\vec{x}) > \theta\}$ depends on the metric d. In the simplest case of Euclidean metric d_{L_2}, the receptive fields of all detectors are hyperspheres in $\mathcal{A}$ centered at $\vec{p}$ and with the same radius determined by θ. Hence, they have the same shape regardless of the distribution of values in their sample sets. This may be undesirable.

In the original theory of conceptual spaces (Gärdenfors, 2000), the use of weighted Euclidean metric is suggested to express unequal importance of dimensions depending on the context or shifts of attention. In our approach, each detector uses *its own* metric derived from the properties of its sample set, instead of a common metric.

Hence, instead of remembering the whole sample set, each detector keeps only a list of dimensions characterizing the subspace $\mathcal{A}$, the prototype of

the category and parameters of the local metric (such as variances or the covariance matrix). We will return to this issue in Section 8.2.3.

Local metrics make similarity judgments essentially asymmetric. In general, for two detectors centered in $\vec{p_1}$ and $\vec{p_2}$, the value of $r_{\vec{p_1}}(\vec{p_2})$ does not have to be equal to $r_{\vec{p_2}}(\vec{p_1})$. People show the same effect, e.g. Tel Aviv is judged more similar to New York than vice versa (Tversky, 1977). Now we review several metrics and their effect on the representational power of the detectors.

8.2.1 Variance-Based Metrics

Euclidean metric weighted by the inverse of the common variance σ^2 of values of all attributes in the sample set

$$d_{L_2,\sigma}(\vec{p},\vec{x}) = \sqrt{\sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - p_i)^2} = \frac{1}{\sigma}\, d_{L_2}(\vec{p},\vec{x})$$

enables representing categories with different levels of generality (hyperspheric receptive fields with radii proportional to σ). Moreover, if we allow infinite weights and define $\infty \cdot 0 = 0$, a category with zero variance will have 1-point receptive field in $\vec{p}$ and will represent an individual.

The natural extension of the previous case is to record variances individually for each dimension. Normalized Euclidean metric

$$d_{L_2,\vec{\sigma}}(\vec{p},\vec{x}) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - p_i)^2}{\sigma_i^2}}$$

with differences on each dimension weighted by the inverse of the variance of sample values on that dimension makes the detector sensitive to the unequal importance of attributes for the category membership. This is very important for cross-situational disambiguation of meanings. The attributes with nearly the same value in all examples will be considered more important for category membership than attributes with big variances within the sample set. The value of an attribute with zero variance will become mandatory for the category instances (any other value in the input frame would yield zero activity of the detector).[4]

The receptive fields of the detectors based on the metric $d_{L_2,\vec{\sigma}}$ are $\vec{p}$-centered n-dimensional hyperellipses having axes of lengths proportional to σ_i. The axes are parallel with those of the input space $\mathcal{A}$ (see Figure 8.2).

[4]For example, triangles can have various sizes, positions, orientations, etc., but they all must have 3 vertices.

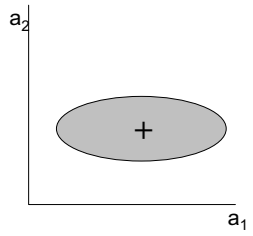

Figure 8.2: Variance-based detectors can account for unequal importance of attributes; the hyperelliptic receptive field has a longer axis along the dimension a_1 because of the bigger variance of a_1 values in the sample.

In case of a mandatory attribute value, the receptive field is a degenerate ellipsoid with the corresponding axis of zero length.

8.2.2 Covariance-Based Metric

The proposed variance-based detectors can learn to attend to differences in some attributes more than in others. However, they cannot learn correlations between attributes, while people can do so (Medin et al., 1982). For example, to induce the concept of *square* from example frames containing the attributes *vertices*, *sizeX*, *sizeY*, one must not only learn the mandatory value 4 of the attribute *vertices*, but also learn that values of attributes *sizeX* and *sizeY* should be identical. This can be achieved by a detector using squared Mahalanobis distance

$$d^2_{\mathbf{\Sigma}^{-1}}(\vec{p},\vec{x}) = (\vec{x}-\vec{p})^\top \mathbf{\Sigma}^{-1}(\vec{x}-\vec{p}) \ ,$$

where $\vec{p}$ and $\vec{x}$ are column vectors and $\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix of the detector's sample set.

The square symmetric $n \times n$ covariance matrix $\mathbf{\Sigma}$ of the sample set

$$\left\{ \vec{x}^{\,(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)} \right) \mid i = \overline{1,N} \right\} \ ,$$

with the mean vector $\vec{p}$ is defined as

$$\begin{aligned} \mathbf{\Sigma} &= (\sigma_{ij})_{i,j=\overline{1,n}}, \text{ where} \\ \sigma_{ij} &= \frac{1}{N-1} \sum_{k=1}^{N} \left(x_i^{(k)} - p_i \right) \left(x_j^{(k)} - p_j \right) \ . \end{aligned}$$

Because $\mathbf{\Sigma}$ is a square symmetric positive semi-definite matrix, it can be decomposed to

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^\top \ , \tag{8.2}$$

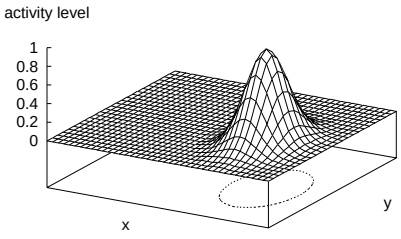

Figure 8.3: A 2-dimensional locally tuned detector with the multivariate Gaussian activity curve. The detector's receptive field with the threshold $\theta = 0.1$ is shown in (x, y) plane.

where $\mathbf{U}$ is an orthonormal rotation matrix $(\vec{e_1}|\vec{e_2}|\ldots|\vec{e_n})$ of eigenvectors[5] of $\mathbf{\Sigma}$ and $\mathbf{D} = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of eigenvalues of $\mathbf{\Sigma}$ with $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ (Ientilucci, 2003). Then computation of the squared Mahalanobis distance

$$\begin{aligned} d^2_{\mathbf{\Sigma}^{-1}}(\vec{p}, \vec{x}) &= (\vec{x} - \vec{p})^\top \mathbf{\Sigma}^{-1} (\vec{x} - \vec{p}) \\ &= \left(\mathbf{U}^\top (\vec{x} - \vec{p})\right)^\top \mathbf{D}^{-1} (\mathbf{U}^\top (\vec{x} - \vec{p})) \ . \end{aligned}$$

can be geometrically interpreted as the standard Euclidean distance of vectors $\vec{x}$ and $\vec{p}$ transformed to a new space with rotated (by $\mathbf{U}^\top$) and scaled (by $\mathbf{D}^{-1}$) dimensions.

Rotation does not change the shape and the size of the receptive field, which are completely determined by the diagonal matrix D. Hence, the receptive field of a detector using the squared Mahalanobis distance will be a hyperellipse with axes of lengths proportional to λ_i and the orientation determined by the rotation matrix $\mathbf{U}^\top$.

By setting $k = \frac{1}{2}$ and using squared Mahalanobis distance in Equation (8.1), we get a detector with (not normalized) multivariate Gaussian tuning/activation curve. The receptive field is a projection of the curve to a hyperplane determined by the threshold θ (see Figure 8.3).

Variance-based metrics are special cases of Mahalanobis metric with diagonal covariance matrices $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and $\mathbf{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_n^2)$, respectively.

[5] $\vec{e_1}, \vec{e_2}, \ldots, \vec{e_n}$ are column vectors.

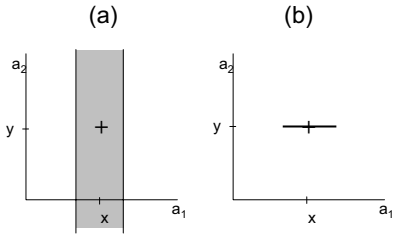

Figure 8.4: The example of a category with the mean (x, y) and the zero variance of the attribute a_2. A detector using the pseudoinverse will consider the attribute a_2 unimportant (a), while a detector using the inverse with infinite values will consider the value $a_2 = y$ mandatory (b).

Singular Case

In case the covariance matrix is singular, hence non-invertible, the Moore-Penrose pseudoinverse $\mathbf{\Sigma}^+$ is often used instead of $\mathbf{\Sigma}^{-1}$. Computation of the pseudoinverse matrix is based on Singular Value Decomposition (SVD) of the matrix $\mathbf{\Sigma}$, which takes the form (8.2). In case of a singular $\mathbf{\Sigma}$,

$$\mathbf{D} = diag(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0) \text{ for some } k < n \ .$$

Then

$$\mathbf{\Sigma}^+ = \mathbf{U}\mathbf{D}^+\mathbf{U}^\top \text{ , where } \mathbf{D}^+ = diag\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_k}, 0, \ldots, 0\right) \ .$$

A detector using the pseudoinverse will ignore the very dimensions that are constant in the whole sample set instead of considering them mandatory – their weights will be zero instead of infinity, and the corresponding axes of the degenerate hyperellipse will have infinite lengths (see Figure 8.4). This is against the philosophy of capturing the regularities of the sample set. For example, the constituting property of the category *triangle* is "having 3 vertices". However, as all the examples of the category have the same number of vertices regardless of other properties, the covariance matrix will be singular and the whole dimension *vertices* will be ignored, because of receiving a zero weight in $\mathbf{D}^+$.

Hence, instead of $\mathbf{D}^+$, we shall use the standard inverse

$$\mathbf{D}^{-1} = diag\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_k}, \infty, \ldots, \infty\right) \tag{8.3}$$

allowing infinite elements $\frac{1}{\lambda_i} = \infty$ for $\lambda_i = 0$. The result of the distance function will be finite, only if the rotated vector $\mathbf{U}(\vec{x}-\vec{p})$ has a zero coordinate on the respective dimensions corresponding to infinite elements of $\mathbf{D}^{-1}$.

However, detectors based on the pseudoinverse can be useful for distinguishing the figure from the background, if examples contain attributes that are constant throughout the sample set, but irrelevant for the category.

Category Generalization Based on SVD-Filtering

After seeing a certain number of examples of some concept, people can decide which properties are relevant for the concept by comparing their variances. For example, if people had to induce the concept *small* from a set of small things of all shapes and colors, they could observe that, albeit finite, variances of shape and color are significantly larger than that of size.[6] Hence, shape and color could be ignored.

In our model, this type of generalization corresponds to finding those diagonal elements of $\mathbf{D}^{-1}$ that are very small in comparison to others, and replacing them with zero. Because $\frac{1}{\lambda_1} \leq \ldots \leq \frac{1}{\lambda_n}$, we can find the largest L such that

$$\frac{\sum_{i=1}^{L} 1/\lambda_i}{\sum_{i=1}^{n} 1/\lambda_i} < b \ , \tag{8.4}$$

where b is some percentage threshold, e.g. $b = 10\%$, and we can set the first L elements of $\mathbf{D}^{-1}$ to zero. The idea is to abstract away those components that contribute little to the total distance. This can be viewed as an opposite process to Principal Component Analysis (PCA), which reduces the dataspace to components with largest variances (Haykin, 1999).

Telling the Figure from the Ground

It follows from (8.4) that if $\mathbf{D}^{-1}$ contains some infinite elements (corresponding to totally invariant properties of the sample set), all finite ones will be abstracted away. This is right for most concepts, but sometimes the infinite elements can be an artifact of taking into account some constant but irrelevant attributes.

This is a common problem of all induction algorithms that only learn from positive examples of a category (Gold, 1967). A property shared by all positive examples should be considered irrelevant, if it is also shared by negative examples. However, our algorithm does not receive and utilize any information about negative examples.

[6] In this simplified example, we abstract away from the problem of semantic dependency of the adjective *small* (see Section 12.1).

8.2.3 Economy of the Representation

Because of the cognitive economy reasons, locally tuned detectors do not record the whole sample sets, but only extract statistical properties necessary for computing the activation function, i.e. dimensions of the subspace $\mathcal{A}$, the prototype $\vec{p}$ and the covariance[7] matrix $\mathbf{\Sigma}$. For technical reasons, the detectors also record the number N of examples seen so far. Hence, we can characterize each locally tuned detector r as a quadruple $r = \langle \mathcal{A}, \vec{p}, \mathbf{\Sigma}, N \rangle$.

As the examples of a category are not usually given all at once, but come sequentially one by one, the mean and the covariance matrix are continuously recomputed by iterative formulas.

Let $N-1$ be the number of examples seen so far, $\vec{x}^{(N)}$ be a novel (N-th) example. Then for $N = 1$,

$$\begin{aligned} \vec{p}^{\,(1)} &= \vec{x}^{\,(1)} \\ \mathbf{\Sigma}^{(1)} &= (0)_{n\times n}, \text{ or } \sigma^2 I_n \ , \end{aligned}$$

where σ^2 is some initial estimate of the variance. For $N > 1$,

$$\begin{aligned} \vec{p}^{\,(N)} &= \frac{N-1}{N}\vec{p}^{\,(N-1)} + \frac{1}{N}\vec{x}^{\,(N)} \\ \mathbf{\Sigma}^{(N)} &= \frac{N-2}{N-1}\mathbf{\Sigma}^{(N-1)} + \\ &\quad + \frac{N}{(N-1)^2}\left(\vec{x}^{\,(N)} - \vec{p}^{\,(N)}\right)\left(\vec{x}^{\,(N)} - \vec{p}^{\,(N)}\right)^{\top} \ . \end{aligned}$$

The attribute set determining the subspace $\mathcal{A}$ is updated iteratively, too, by intersecting with the attribute set of each new example. If some attributes are removed from the subspace $\mathcal{A}$ this way, the corresponding rows and columns of the covariance matrix and the mean vector are removed, too.

8.2.4 Sign Pattern Based Detectors

In everyday reasoning, people often abstract away from numerical values and use a simpler qualitative calculus based on ordinal relations and invariant sign patterns (Kuipers, 1994; Takáč, 1997, 2003c). The sign structure of attributes is often constitutive for relational concepts, e.g. the relation $bigger(f_x, f_y)$ can be expressed as $f_x.size > f_y.size$, or equivalently, $\mathrm{sgn}(f_x.size - f_y.size) = +1$.

[7] As we have already mentioned, variance-based metrics are special cases of the covariance-based one.

Now we define a qualitative detector sensitive to the sign pattern of attributes in the sample set. The subspace $\mathcal{A}$ is defined by a set of attributes present in all examples with the same sign. The projection of frames to $\mathcal{A}$ is composed with the operator sgn. The sign pattern is recorded in the prototype

$$\vec{p} = (p_1, \ldots, p_n), \text{ where } p_j = \operatorname{sgn}\left(x_j^{(i)}\right) \ \forall\, i = \overline{1, N} \ .$$

The sign pattern is recorded only once, upon seeing the first example. Later updates of the criterion only remove from $\mathcal{A}$ the attributes not occurring in new examples with the same sign as recorded. The detector returns a binary result: 1, if an input frame has the same sign pattern as $\vec{p}$ for all attributes in $\mathcal{A}$, and 0 otherwise.[8]

8.3 Identification Criteria Based on Locally Tuned Detectors

8.3.1 Object Criteria

Object criteria operate on single frames. The object criteria can represent individual objects, if they return zero for all but one particular frame, e.g. *JohnSmith*(f), properties of objects, e.g. *married*(f), *large*(f) or *credible*(f), and classes of objects, e.g. *student*(f), *fruit*(f), *desktopComputer*(f).[9] Actually, there is no formal difference between criteria of properties and classes.

Categories of individual objects, classes and properties of objects can be directly represented by locally tuned detectors (as their argument is one frame describing an object). Criteria having more input arguments (relations, changes) can be reduced to locally tuned detectors by transforming their input.

8.3.2 Relational Criteria

Binary relational criteria, e.g. *larger*(f_1, f_2) or *near*(f_x, f_y) for frames f_1 and f_2, can be represented by (quantitative or qualitative) detectors operating on a transformed input $\Delta(f_1, f_2)$, where $\Delta(f_1, f_2) = f$ is a frame of differences of aligned attributes (Markman and Gentner, 1993), defined by

$$A_f = A_{f_1} \cap A_{f_2} \text{ and } f.a = f_1.a - f_2.a \quad \forall\, a \in A_f \ .$$

[8]The sign pattern based detector can be viewed as a special case of the variance-based detector with all attributes mandatory (as their signs have zero variance).

[9]Mnemonic identifiers of criteria should not be confused with language expressions (words), which differ by font and quotation marks, e.g. *large* vs. "*large*".

8.3.3 Criteria of Situations

Complex situations or properties of the whole scenes, e.g. concepts of a *risky-Investment* or a *catOnHotTinRoof*, can be built as hierarchical networks of locally tuned detectors. Detectors of the bottom level operating on perceptual frames represent components of the situation and their required mutual relations. Vectors of output activities of the elementary detectors serve as an input to aggregate detectors of the higher level, which can attribute unequal importance to the elementary detectors and/or detect their mutual correlations.

In an aggregate detector characterized by $\langle \mathcal{A}, \vec{p}, \mathbf{\Sigma}, N \rangle$, the set $\mathcal{A}$ contains the list of aggregated lower-level detectors, the vector $\vec{p}$ represents a prototypical pattern of their activities and $\mathbf{\Sigma}$ determines their importance and required inter-correlations.

As a situation criterion has only one input argument – the whole scene $S^{(t)}$, it is important to specify what perceptual frames should be an input to what elementary detectors of the criterion.

If a situation criterion requires the presence of k objects on the scene and describe their properties and relations among them, we say that it has the arity k. The input argument relations can then be described by a directed multigraph[10] with k vertices corresponding to slots[11] for k objects required on the scene (see Figure 8.5). Each edge between some vertices (i, j), labeled with a lower-level detector r, expresses a required relation $r(\#i, \#j)$ between objects filled in the slots i, j. The order of arguments is important. Loops (i, i) express required properties, i.e. object criteria $r(\#i)$, for an object filled in the slot i.

Formally, a situation criterion with one aggregate detector $s = \langle \mathcal{A}, \vec{p}, \mathbf{\Sigma}, N \rangle$ is characterized by a tuple $\langle s, k, g \rangle$, where k is the arity and g is the argument descriptor $g : \mathcal{A} \to \{1, \ldots k\} \times \{1, \ldots k\}$. I.e., the argument descriptor g determines input arguments for each lower-level detector $r \in \mathcal{A}$.

[10] A (directed) multigraph is a graph allowing multiple (directed) edges between a pair of nodes. We also allow loops, i.e. edges starting and ending in the same node.

[11] The situation criterion is evaluated for all ordered subsets of $S^{(t)}$ with k elements and returns maximum of the results for the k-subsets.

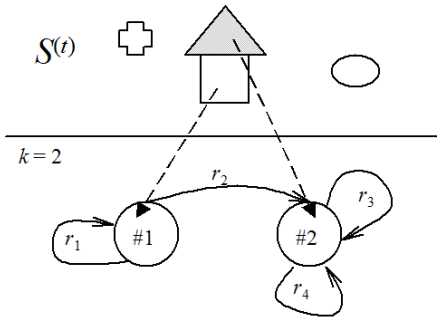

Figure 8.5: An example of a situation criterion for the concept of a house with a grey roof (in a simplified 2D block world). The situation (above the solid line) contains 4 objects, two of which are instantiated (dashed lines) in slots of the binary situation criterion (under the solid line). The criterion $r_1 \rightarrow (1,1)$ represents the fact *square*(#1), the criterion $r_2 \rightarrow (1,2)$ represents the fact *on*(#1,#2), the criterion $r_3 \rightarrow (2,2)$ represents the fact *triangle*(#2), and the criterion $r_4 \rightarrow (2,2)$ represents the fact *grey*(#2).

8.4 Representation of Environmental Dynamics

8.4.1 Change Criteria

Criteria expressing changes of properties of an object in time are relational criteria applied across time to frames $f^{(t)}, f^{(t-1)}$ of the same object. They can be represented by detectors applied to a transformed input $\Delta f_{t,t-1} = \Delta\left(f^{(t)}, f^{(t-1)}\right)$ (see Section 8.3.2).

Environmental dynamics expressed by the change criteria is an important part of meanings of verbs. Some changes can be captured by qualitative relations, e.g. *grow(f)* can be expressed by $\text{sgn}\left(\Delta f_{t,t-1}.size\right) = +1$, others require encoding of a typical change's magnitude, e.g. movement criteria for *crawl*, *walk*, *run* could differ in mean values of $\Delta f_{t,t-1}.position$. The criteria with zero sign pattern of some attributes can represent a state or a persistence of a property, e.g. *stay*.

8.4.2 Criteria of Events

Multiple co-occuring changes of a complex dynamic scene can be represented by criteria of events. Event criteria are a generalized version of criteria of

situations in that their elementary criteria can include criteria of changes.[12]

8.4.3 Verb Semantics

Of course, this is just a part of the picture. Embodied verb representation is connected to actions and includes the manner of performance, e.g. the representation of *jumping* can refer to a non-declarative procedural representation of an invariant motor stereotype together with a frame representing variable parameters of the action, e.g. the velocity or joint angles (Bailey et al., 1997). Other possibilities of representation of verb meanings are mentioned in Section 4.5.

As the embodied meanings are grounded in sensorimotor interactions with the environment, they also include situated causal knowledge about preconditions of successful actions and their possible consequences.[13] For example, the action of *lifting* performed in the same manner (with the same force) can lead to different outcomes (changes) depending on the object of the action, e.g. lifting a ball or lifting a 200 kg piece of furniture. Such propositional knowledge can be suitably represented by cross-categorical associations of the type

$$(preconditions, action \rightarrow consequence) \ ,$$

where *preconditions* are criteria for objects of the action, *action* is a criterion representing the action's manner, and *consequences* are change criteria of the resulting dynamics.

[12] In the argument descriptor multigraph, change criteria determining relations between current and past frames of an object $\#i$ are placed on loops (i, i).

[13] Hence, the representation of complex meanings is not purely *categorical*, but also includes *propositional* elements.

Chapter 9

Individual Construction of Meanings

In this chapter, we describe a novel computational model of construction of environmentally and ecologically relevant meanings, which is another major contribution of this thesis. In the previous chapter, we introduced representation of meanings based on locally tuned detectors that can be constructed from sets of category examples. However, in the course of time, the agent perceives a mixed sequence of instances of many concepts and it must somehow determine which of the existing criteria should be updated by which examples (and when to create a new criterion).

If the agent has no additional information, it can group similar frames to categories by unsupervised clustering techniques, e.g. *Distributed Clustering Algorithm* (Hulth and Grenholm, 1998), *1-nearest neighbor*, or others (Everitt et al., 2001), maximizing inter-cluster and minimizing intra-cluster distances.

The simplest unsupervised learning procedure works as follows:

1. For an input percept[1] f and a set of criteria C, find $r^* \in C$ such that $\forall r \in C : \; r^*(f) \geq r(f)$.

2. If $r^*(f) > \theta$, update r^* by f, else create a new criterion r^{new} with f as its first example (θ is a threshold parameter).

This algorithm clusters the input by its distributional properties. In this way, it can arrive at *environmentally* relevant categorization (i.e. the one that takes into account distribution of properties in the environment).

Construction of *ecologically* relevant categories that take into account causal relations in dynamic environment and are of some use for the agent,

[1]The formulation for pairs of input percepts and for scenes is analogical.

must be based on pragmatic criteria. Now we review our experiment focusing on construction of categories by sensorimotor interactions with the environment (Takáč, 2006a).

9.1 Model

9.1.1 Environment

In the experiment, the simulated environment (non-toroidal 2-dimensional lattice 25×25) contained the agent and 30 other objects – 10 "fruits", 10 "toys" and 10 "pieces of furniture" placed on random positions of the lattice.

The initial values of object attributes were randomly generated as uniformly chosen integers from respective intervals of the pattern $\{$*weight*: 20, *age*: 3, *posX*: [0, 24], *posY*: [0, 24], *posZ*: 0$\}$ for the agent, $\{$*weight*: [1, 3], *size*: [1, 49], *color*: [0, 4], *roundness*: [0, 9], *posX*: [0, 24], *posY*: [0, 24], *posZ*: 0$\}$ for fruits, $\{$*weight*: [1, 9], *color*: [0, 9], *cries*: [0, 1], *dressed*: [0, 1], *posX*: [0, 24], *posY*: [0, 24], *posZ*: 0$\}$ for toys, and $\{$*weight*: [20, 49], *size*: [20, 49], *legs*: [0, 4], *material*: [0, 9], *posX*: [0, 24], *posY*: [0, 24], *posZ*: 0$\}$ for pieces of furniture.

9.1.2 Agent

The agent was actively exploring its environment. In each time step, it could randomly choose an action from its action repertoire (lift, put down) and perform it with different parameters (force, arm angle) upon an object randomly chosen from its surrounding. The action was randomly generated from the pattern $\langle$*actionType*: *liftUp*, $\{$*armAngle*: [1, 9], *force*: [1, 19]$\}\rangle$ or $\langle$*actionType*: *putDown*, $\{$*armAngle*: [1, 9]$\}\rangle$.

The action type modeled an invariant procedural representation of the action (motor stereotype), while the parameter frame represented variable parameters of the action execution.

The effects of the action on the chosen object were simulated by the environment. In the case of lifting (*liftUp* action), the vertical position (*posZ* attribute) of the object was increased by the value proportional to the arm angle (*armAngle*), if the exerted *force* was greater than the *weight* of the object, otherwise the action had no effect. In the case of putting down (*putDown* action), the vertical position *posZ* of the object was decreased by the value proportional to the arm angle, unless the object was already on the ground ($posZ \leftarrow \max(0, posZ - armAngle)$).

9.1.3 Representation

By interacting with the environment, the agent gradually learns to distinguish between object, action and change categories. Object categories are represented by object identification criteria (Section 8.3.1) and actions by *action identification criteria*. These criteria are similar to object criteria, except that they also store a type of the action. They are applied to perceptual frames with action type and execution parameters, and they return zero if the two action types do not match (if they do, the parameter frame is evaluated in the standard way). Change categories are represented by change identification criteria (Section 8.4.1).

Simple Categories

Categories of each type are represented by identification criteria with variance-based detectors, stored in three separate categorical systems C_o, C_c, C_a. This is cognitively plausible, as similar separate representational systems exist in humans (Ungerleider and Mishkin, 1982; Orban et al., 1995; Rizolatti et al., 1996), and the representations remain perceptual (Barsalou, 1999).

Complex Categories

Causal relations between actions, objects and changes are represented in the form of associations among their respective categories (analogy to associative areas of the cortex). Formally the agent's association system V is a set of triples

$$V = \{\langle r_a, r_o, r_c\rangle \mid r_a \in C_a, r_o \in C_o, r_c \in C_c\}\,.$$

9.1.4 Learning Algorithm

The agent in our model learns by observing consequences of its own actions.[2]

Objects and actions are grouped to categories by the change. That is, if an action leads to the same change on several objects, they will all fall in the same category and vice versa. A significantly different outcome of the action triggers creation of new categories.

All action categories associated with some object category represent agent's knowledge of *affordances* (see Section 2.5.2) of the object, while all object categories associated with an action category form the precursor of a verb-centered semantic representation – a *verb island* (Tomasello, 1992).

[2]The agent 'assumes' that all observed changes of the chosen object were caused by its action. Children have the similar attitude, called "magical causality" (Piaget and Inhelder, 1966), even short after their sensory-motor stage.

Initially, the agent starts with the empty category sets C_o, C_c, C_a and the empty association set V. These sets can be modified in each time step, after the agent perceives a triple $\langle f_a, f_o, \Delta f_o \rangle$ of percepts of a performed action, object and its change,[3] in the following way:

1. Find in V the most similar[4] association $v^* = \langle r_a, r_o, r_c \rangle$ to the input triple $\langle f_a, f_o, \Delta f_o \rangle$. (If there is no association with non-zero similarity, create a new one either by reusing existing categories, if they are individually similar enough to the percepts, or by creating new categories.)

2. If $sim(r_c, \Delta f_o) > \theta(t)$, update r_a by f_a, r_o by f_o and r_c by Δf_o, otherwise:

3. if $r_o(f_o) > r_a(f_a)$, create a new action category from f_a, else create a new object category from f_o and use it to form a new association.

In step 2, if the change category of the association is similar enough to the perceived change, the percepts are considered to be the instances of the associated categories and all three categories are updated by the percepts. Otherwise, a new category is created for the less similar percept of either the object, or the action (step 3).

The prediction threshold $\theta(t)$ determines the precision of the representation. It can be constant during the whole simulation or it can increase in time to model the child's growing ability to distinguish differences in the environment.

Merging

It can happen that some categories, which started independently, become very similar after being updated by more examples. In our model, such similar categories are sought and merged in the following way: if any of the systems C_o, C_c, C_a contains two criteria r_1, r_2 with mean cases f_1, f_2, such that $\min(sim(r_1, f_2), sim(r_2, f_1)) > 0.9$, they are replaced by a new criterion r with $A_r = A_{r_1} \cap A_{r_2}$. The means and variances of attributes of r are computed from those of r_1 and r_2 as if they were characteristics of the union of example sets of the original criteria. The merging can propagate to the association level – if, after merging some categories, there exist associations $\langle r_a, r_o, r_{c_1} \rangle$ and $\langle r_a, r_o, r_{c_2} \rangle$, such that $r_{c_1} \neq r_{c_2}$, r_{c_1} and r_{c_2} are merged.

[3] $\Delta f_o = \Delta\left(f_o^{(t)}, f_o^{(t-1)}\right)$, see Section 8.4.1.

[4] The similarity is computed using equation (8.1) from the weighted sum of distances $dist\left(\langle r_a, r_o, r_c \rangle, \langle f_a, f_o, \Delta f_o \rangle\right) = w_a d(r_a, f_a) + w_o d(r_o, f_o) + w_c d(r_c, \Delta f_o)$. However, the overall similarity is zero, if r_a, f_a do not have the same action type.

9.2 Measures and Parameters of Model Simulations

In each time step after performing an action, the agent adapted its representation according to the association algorithm described above. To evaluate the usefulness and adequateness of the representation, we measured its ability to predict the correct result of the action. After choosing an object and an action, the agent found the association with the highest similarity of object and action.[5] The change criterion of that association was then applied to the perceived change and the resulting activity was recorded as the *prediction*. However, the more general change criteria give higher similarity values, therefore, we also measured the *generality* of the prediction expressed by the average standard deviation of attributes of the criterion used for prediction (lower value means higher accuracy of the prediction). We also measured the number of criteria in the agent's representation. Each measure has been averaged over the time window of 20 last steps.

The parameters of the association algorithm were $w_o = 1$, $w_a = 100$, $w_c = 1000$. In order to model developmentally growing sensitivity to environmental differences, we used detectors with increasing decision threshold θ for change criteria. The threshold $\theta(t)$ was linearly increasing from $\theta(0) = 0$ to $\theta(700) = 0.7$ and constantly equal to 0.7 for $t > 700$. Results of the experiments were averaged over 30 simulation runs.

9.3 Results

Before we present the results, we review that we simulated the model of category formation by sensorimotor interactions with objects in the environment. The agent performed random actions on randomly chosen objects and grouped objects, actions and changes into categories according to the result of interaction. The goals of the experiment were: (1) verify that the proposed mechanism can lead to construction of ecologically relevant categories, (2) evaluate the effects of category merging, (3) compare the effectiveness of the representation based on identification criteria with prototypes in conceptual spaces.

[5] Computed from the weighted distances of object and action, see the footnote 4.

9.3.1 General Results

In the first experiment, the agent did not use merging. We let the agent interact with its environment for 5000 time steps.

The results are summarized in Figure 9.1a. As we can see in the graph, while the prediction threshold is low, the agent only uses a few basic criteria. After the threshold rises over a certain value (around 0.5), the number of criteria starts to rapidly increase, which leads to a better accuracy of the prediction. As the threshold stabilizes at the value of 0.7, the total number of criteria slowly saturates, together with the generality exponentially decaying to a certain value. The prediction value converges to approximately 0.7. Recall that this value corresponds to the average distance σ, which is an average intra-cluster distance of the category. Hence, this means that the criteria give correct predictions and the agent has constructed ecologically relevant categories.

9.3.2 Merging

In the second experiment (Figure 9.1b), the agent merges similar criteria every 50th time step since the time 1000. This decreases the total number of the criteria at the cost of more general predictions (the generality does not exponentially decay, but stays between 1 and 1.5). The prediction value again converges to approximately 0.7, i.e. the criteria give correct predictions.

9.3.3 Comparison to Prototypes

The advantage of the variance-based identification criteria is that they are sensitive to unequal importance of attributes (or scaling of different dimensions) for category membership. In order to compare them with a standard prototype representation, we ran an experiment with criteria based on standard Euclidean metric (Figure 9.1c). Despite that the criteria were merged as in the second experiment, the number of criteria is almost double and the prediction value is lower than in the case with variances. This means that the prototype representation is less compact and less effective for predictions.

9.3.4 Representation in Detail

Actual meanings of the constructed categories can be guessed by inspecting the internal representation of the agent. In Table 9.1, we can see a fragment of the agent's representation from an example run of the experiment, in which the agent acquired 4 object criteria, 7 action criteria, 10 change criteria and

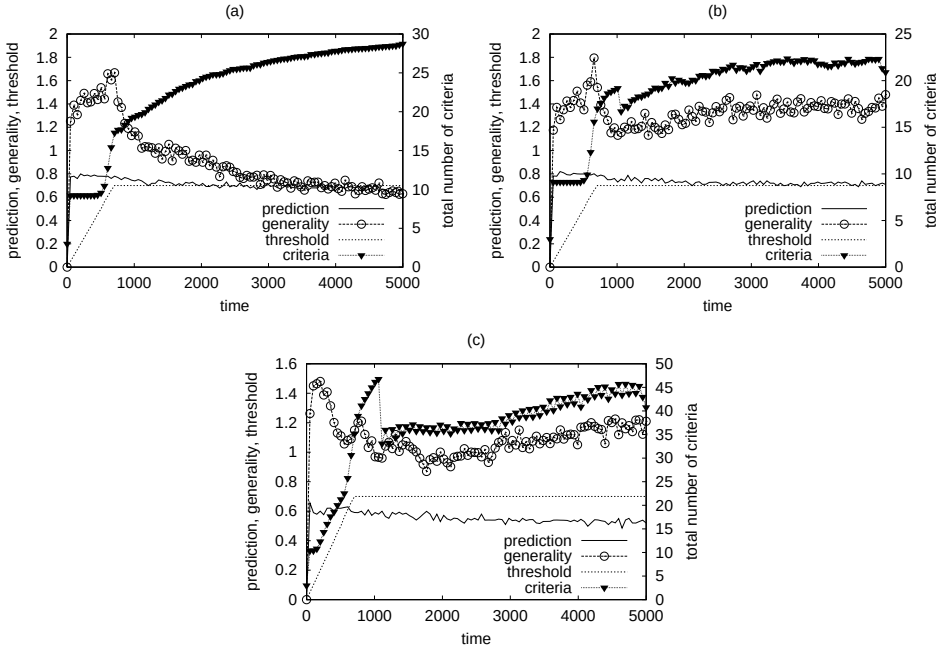

Figure 9.1: Construction of categories based on outcomes of sensorimotor interactions. The measure of *prediction* expresses the agreement of the predicted outcome (change) of the performed action with its real outcome, the *generality* express the inaccuracy (permissible deviation) of the predicted change, the *threshold* is a developmental parameter that determines when two changes are considered the same. *Criteria* is the total number of criteria in the representational system of the agent. **(a)** Experiment with no merging of criteria. While the prediction threshold is low, the agent only uses a few basic criteria. Then the number of criteria starts to rapidly increase, which leads to a better accuracy of the prediction. As the threshold stabilizes, the total number of criteria saturates and the generality decays to a certain value. The value of prediction converges to that of average intra-category distance (0.7), i.e. the predictions are correct and the agent has constructed ecologically relevant categories. **(b)** The effect of criteria merging: it keeps the number of criteria lower than in the experiment (a), at the cost of lower accuracy (higher generality). **(c)** Comparison to prototypes (criteria insensitive to variances of attributes): Agents used much more criteria than in the experiment (a) and still achieved a lower prediction value.

Table 9.1: Example of object criteria (above) and a fragment of associations (below) constructed by the agent interacting with its environment by sensation and action. For brevity, the attribute values are written as the mean (prototype) $\pm$ the standard deviation σ. In the table below, object categories are in rows, action categories are in columns, associated change criteria are in intersections of rows and columns. Arguments of actions are $putDown(armAngle)$ and $liftUp(armAngle, force)$. Change criteria express the difference in vertical position of the involved object.

	posX	*posY*	*posZ*	*weight*	*color*
C1	13 ± 7	13 ± 8	0 ± 0	37 ± 23	
C2	14 ± 7	13 ± 8	4 ± 13	39 ± 22	
C3	11 ± 6	10 ± 7	35 ± 28	4 ± 3	2 ± 2
C4	11 ± 6	10 ± 7	25 ± 23	4 ± 3	

	Action	
Category	$putDown(5 \pm 3)$	$liftUp(6 \pm 2, 10 \pm 6)$
C1	no change	
C2		no change
C3	Δ={*posZ*: -6 ± 1}	Δ={*posZ*: 7 ± 1}
C4	Δ={*posZ*: -4 ± 2}	Δ={*posZ*: 5 ± 2}

formed 13 associations. Table 9.2 shows the object criteria applied to 31 objects in the environment. Numbers in a row express the object counts of a given type most similar to the criterion in a column. We can see that the agent constructed categories such as "objects too heavy to be lifted" (C2) or "objects that cannot be put down, because they are already on the ground" (C1). Category C3 represents mostly fruits and C4 mostly toys. As attributes other than *weight* or *posZ* are present in the criteria too, they could help the agent in classification (e.g. if all heavy objects were in the same part of the grid, or had some specific color). Hence, the representation is situated and it encodes the learning context.

9.3.5 Discussion

The results of the experiments show that the proposed mechanism of action-based category formation leads to ecologically relevant categories, i.e. such

Table 9.2: Construction of categories by sensorimotor interactions. Number of objects of each type for a constructed category they are most similar to.

	Category			
Object type	C1	C2	C3	C4
agent	1			
fruit			8	2
toy		1	3	6
furniture	5	5		

that support prediction of results of the agent's own actions. Once an agent can represent predictions about the outcome of it's actions, it can use them for planning sequences of actions to satisfy it's needs and goals.

The experiments with category merging show the trade off between usability of the representation and memory load: lower number of more general categories versus higher number of specialized categories that support more accurate predictions.

The last experiment demonstrates that the identification criteria based representation is more effective and compact than the representation based on prototypes in conceptual spaces due to its sensitivity to unequal importance of attributes for category membership.

Chapter 10

Construction of Meanings by Social Instruction

The experiment described in the previous chapter modeled individual construction of preverbal meanings. In this chapter, we describe a computational model focusing on the study of the influence of verbal instruction (naming) on category formation process, in line with empirical observations of Waxman (2004) described in detail in Section 3.3.2 (we review that their observations suggest that consistent using of the same name for distinct objects motivates the infant to look for *similarities* and promotes formation of categories). The model can also be viewed as a test of the weak version of the Sapir-Whorf hypothesis (Whorf, 1956) stating that language affects our conceptual system. The proposed computational model is however our original contribution.

10.1 Model

The model consisted of two agents situated in a simulated environment: a teacher describing various aspects of the present situation, and a learner inducing meanings of the teacher's words by noticing cross-situation similarities between their referents.

10.1.1 Environment

The simulated environment consisted of 2D geometrical shapes characterized by five attributes: the number of vertices ($vertices \in [2, 5]$), coordinates of the centroid of the shape ($posX, posY \in [0, 50]$) and the size of the bounding rectangle ($sizeX, sizeY \in [0, 25]$). The initial values of attributes were uniformly randomly generated integer numbers from the respective intervals.

The environment was dynamic in that in each time step, randomly selected objects could be resized, moved, or removed from the environment and newly generated objects could be added (there were usually 2-4 objects simultaneously present on the scene). Multiple changes could happen simultaneously in one time step.

10.1.2 Learning Mechanism

In each time step t, both agents (the teacher and the learner) perceived the same scene $S^{(t)}$. The teacher produced the scene description $D_\theta(S^{(t)})$ using its own criteria. This description composed of words and their referents (identified by foci, see Section 7.5.1) served as the learner's learning input.

The learner constructed its identification criteria from examples determined by the foci and associated them with the received words. A meaning of each word was induced from all the referents that the word has been used with.

The induction of meanings was guided by *no true synonymy* and *no true homonymy* assumptions. Although natural languages do contain words with multiple meanings (homonyms) and multiple expressions for a single meaning (synonyms), in case of bootstrapping the language and concepts from scratch, it is useful to start with no homonymy and no synonymy.[1]

1. *No true synonymy:* Different words have different meanings, even if they share a referent (in that case they express different aspects of the referent).[2] Put in practice, if the agent hears an unknown word w in the context of a referent $\phi(S^{(t)})$, a new criterion r is created with $\phi(S^{(t)})$ being its first example and r is associated with w in the agent's lexicon.

2. *No true homonymy:* A single word has a single meaning, even if it is used with more referents. This assumption is crucial for cross-situational disambiguation of the meaning: all the referents of a single word across different situations are considered instances of the same category denoted by the word.[3] Put in practice, if the agent already knows some meaning r of a word w and w is now used with a new referent $\phi(S^{(t)})$, r is updated by $\phi(S^{(t)})$ (instead of creating a new criterion).

[1] Children acquiring a language use similar constraints, see Section 3.3.3.

[2] This assumption corresponds to the *Principle of Contrast* (Clark, 1987).

[3] If a word has been used in apparently different contexts (e.g. if the referents have nothing in common), the agent can detect homonymy and associate multiple criteria with the word. However, this feature has not been implemented in our model.

Example.

Let us consider an agent that lives in a world of geometric shapes placed on a 50×50 grid with the point coordinates $(1, 1)$ on the left bottom and $(50, 50)$ on the right top. If the agent perceives an object

```
f = {vertices: 3; size: 18; color: 3; posX: 1; posY:23}
```

denoted by words "*left*", "*big*", "*triangle*", it creates three identification criteria, which are initially identical and represent the "snapshot" of the perceived object f. The criteria begin to differentiate, when they are updated by more and more instances. E.g. the "snapshot" criterion associated with the word "*triangle*" will be updated by frames of various objects having all kinds of colors, positions, sizes and other properties, but all having 3 vertices. Attributes not common to all instances will be removed from the criterion and others will gain lower importance because of their high variance in the sample. Hence, the property of having 3 vertices (with zero variance in the sample) will become decisive in the criterion associated with the word "*triangle*". Also the word "*left*" will be heard with many different objects sharing the property of low value of the attribute *posX*, etc. The more contexts of the word's use, the bigger the probability that the referents will vary in the properties irrelevant for the meaning of the word. However, if e.g. all triangles in the agent's world are big, then having a big size will become part of the meaning of the word "*triangle*". Hence, the induced representation is situated and contextual.

10.2 Measures and Parameters of Model Simulations

The experiment was run for 5000 learning epochs (time steps). In each time step, the teacher, using a predefined ontology and the lexicon, described the current scene (including the changes) to the learner. The teacher's lexicon included 2 nouns, 3 adjectives and 2 verbs (see Table 10.1).

The learner used detectors based on the Mahalanobis metric, with the receptive field threshold $\theta = 0.1$ and SVD-filtering with the threshold $b = 10\%$ (see Section 8.2.2).

To evaluate the fidelity of meaning transmission, before receiving the scene description from the teacher, the learner described a scene in each time step too, and the two descriptions were compared for the *correctness* and the *completeness* of the learner's description. The measure of the *description similarity* was the average of the correctness and the completeness.

Table 10.1: A predefined ontology and lexicon of the teacher in the experiment focusing on the influence of naming on category formation process.

Word	Meaning
square	$vertices = 4 \wedge sizeX = sizeY$
triangle	$vertices = 3$
big	$sizeX > 15 \wedge sizeY > 15$
slim	$sizeX < 0.2 sizeY$
small	$sizeX < 10 \wedge sizeY < 10$
grow	$sizeX^{(t)} > sizeX^{(t-1)} \wedge sizeY^{(t)} > sizeY^{(t-1)}$
shrink	$sizeX^{(t)} < sizeX^{(t-1)} \wedge sizeY^{(t)} < sizeY^{(t-1)}$

The *correctness* of the learner's description was computed as $1 - w/L$, where w was the number of wrong words in the learner's description of the scene and L was the total number of words in the learner's description. A word in a learner's utterance describing some referent was considered *wrong*, if it was not used by the teacher in its utterance describing the referent.

The *completeness* of the learner's description was computed as $1 - m/T$, where m was the number of teacher's words missing in the learner's description of the scene and T was the total number of words in the teacher's description. A word in a teacher's utterance describing some referent was considered *missing*, if it was not used by the learner in its utterance describing the referent.

We also evaluated a pragmatic *usage* of the learner's ontology in guessing games. In each time step, the teacher uttered a verbal description of a referent randomly picked up from the scene and the learner guessed the referent. The learner's guess was a set $\mathcal{L}$ of possible referents of the utterance, as understood by the learner. In case $\mathcal{L}$ did not contain the referent meant by the teacher, the usage was zero. Otherwise, the success in the guessing game was evaluated by comparing $\mathcal{L}$ with the set $\mathcal{T}$ of referents that the teacher itself would guess from the utterance (as the agents did not play discrimination games in our model, the teacher's description did not have to be unique either). The usage was computed as $1/(1 + r)$, where r was the number of referents in $\mathcal{L} - \mathcal{T}$. Hence, even in the case of a correct guess, the usage was lowered by any extra referents that could not be meant by the teacher.

The *uncertainty* inherent in the teacher's descriptions was measured as $1 - 1/|\mathcal{T}|$. For example, if the teacher described the chosen object by the utterance "triangle", and the scene contained two objects categorized as triangles by the teacher, the uncertainty of the teacher's description would be

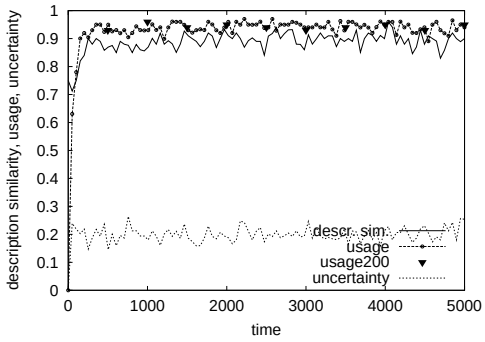

Figure 10.1: Cross-situational learning of categories from verbal instruction. The quality of the lexicon acquired within one generation. The *description similarity* expresses the similarity of verbal descriptions of the current scene produced by the teacher and the learner. The *usage* is the measure of success in guessing games played during the learning; the *usage200* is the average success in 200 guessing games played after each 500 learning epochs. The *uncertainty* is the degree of ambiguity inherent in the teacher's description. Each measure in the graph has been averaged over the time window of 30 last steps and the results of the experiment were averaged over 10 simulation runs with different random seeds. Cross-situational learning is fast and reliable: the learner soon acquires an ontology and lexicon of a sufficient quality, which remains stable for the rest of the simulation.

50%. If the teacher's utterance had a unique referent, uncertainty would be zero.

Besides playing guessing games during the learning, the agents played 200 guessing games after every 500 learning epochs. The measure "usage200" is an average usage of 200 guessing games played after learning. The guessing games were only played for evaluation purposes and did not have any influence on the learning process.

10.3 Results

The goal of this experiment was to validate the proposed mechanism of meaning construction based on the influence of naming on category formation process. This goal has been satisfied; the simulation results (Figure 10.1) show that cross-situational learning is fast and reliable: the learner soon acquires an ontology and lexicon of a sufficient quality, which remains stable for the rest of the simulation. We discuss these results in more detail in Chapter 12.

Chapter 11

Meanings in Intergenerational Transmission

We have shown how a learner can construct meanings sufficiently similar to those of its teacher by verbal instruction. Now the question is, whether meanings constructed this way remain stable, if we let the acquisition process iterate intergenerationally. In this chapter, we describe an extension of the previous experiment, based on the *iterated learning model* (ILM, see Section 4.2.2) framework (Kirby and Hurford, 2001). Our experiment is designed to study how *meanings* change across generations (Takáč, 2007c,a).

The iterated learning model, which involves vertical cultural transmission of language between generations, has primarily been designed for modeling the emergence of grammar. In the ILM framework, language develops by flowing between two forms of private language competence and externalized utterances by processes of acquisition and production. A learner builds up its own internal language representation by observing external language input from its teacher, later the learner becomes a teacher and produces utterances, which are the input for the next generation learner, etc.

11.1 Model

The first generation setting in the model was identical to the experiment described in the previous chapter, except that we varied the number of learning epochs. After a certain number of epochs, the teacher with predefined ontology was removed and the learner became a teacher for a new agent with an empty ontology and lexicon. We let this process iterate for 50 generations.

We ran two versions of the experiment: in Experiment 1, the agent could neither modify nor add any new meanings, once it became a teacher (it only

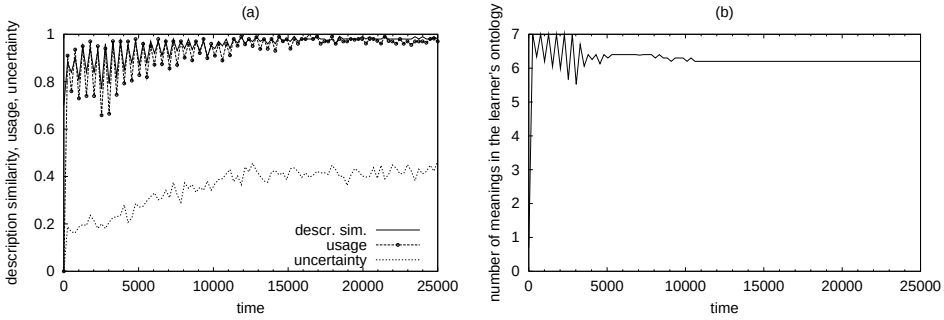

Figure 11.1: Iterated intergenerational transmission of meanings by verbal instruction. Generations exchanged every 500 time steps. The results were averaged over 10 simulation runs with different random seeds (for description of measured parameters, see Section 10.2). **(a)** High values of description similarity and the pragmatic usage were quickly retained after each drop caused by the generation exchange. They gradually stabilized at a very high value close to 1, at the cost of a higher uncertainty caused by overgeneralization of some meanings. **(b)** The average number of meanings stabilized on the value of 6 out of 7 original meanings of the first-generation teacher (one meaning died out because of overspecialization).

used the meanings acquired from it's own teacher). In Experiment 2, the teacher could invent new meanings or extend old ones, in case it had no meanings applicable to describe some object on the current scene.

11.2 Experiment 1

Figure 11.1 shows the results of Experiment 1 run for 50 iterations of 500 learning epochs. We measured the description similarity and the pragmatic usage of the learner's lexicon in each generation. As we can see, the quality of the lexicon is quickly retained after each drop caused by a generation exchange.

To explore inter-generational meaning shifts, we also inspected the agents' internal representations of categories. We found out that categories *grow* and *shrink* represented by sign pattern based detectors remained the same in all generations. The agents were also successful in inducing the correct criteria for *triangle* (3 vertices, all other attributes irrelevant) and *square* (4 vertices and equal side lengths), which remained the same in all generations, too. Though the exact parameters of the criterion for *slim* varied across generations, the property of having horizontal size small in comparison to the vertical one has been correctly captured and retained. Criteria for *big*

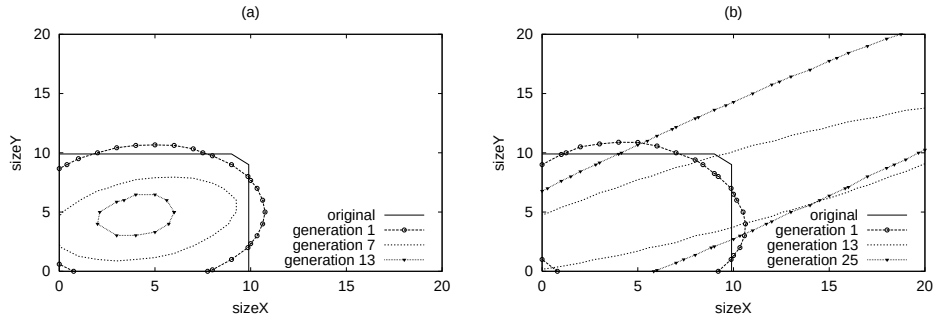

Figure 11.2: Overspecialization and overgeneralization – two sources of instability in iterated intergenerational meaning transmission. The receptive field of each displayed category was projected into the plane with dimensions $sizeX, sizeY$. The projecting plane crossed other dimensions in mean values of the dimensions recorded in the category's detector. In some simulation runs, the categories *big* and *small* showed one of the following behaviors: **(a)** overspecialization – the size of the receptive field converged to zero over generations, **(b)** overgeneralization – a random correlation of some attributes in the sample overtook other attributes that became overgeneralized (ignored).

and *small*, based on certain intervals of uncorrelated attribute values, did not turn out to be so stable. Either they were overspecialized in some simulation runs and they died out (their receptive field gradually shrank to zero), or they were overgeneralized (due to a takeover of some attributes, see Figure 11.2).

In order to explore causes of this instability, we varied the number of learning epochs in each generation (Figure 11.3). We can see that a smaller number of learning epochs causes smaller sample sets. If sample sets are too small, concepts are unstable, some of them get overspecialized and disappear (the number of total meanings gets smaller), others get overgeneralized (the uncertainty rises). This is the case of simulations with less than 300 learning epochs per generation (corresponding to less than 25 examples for the detector with the smallest sample set). In simulations with more than 300 learning epochs, the average number of meanings stays between 6 and 7 and the uncertainty is around $30 - 35\%$. These results are further discussed in the following section.

11.2.1 The Influence of the Meaning Bottleneck

As objects and their changes are generated randomly within the fixed number of learning epochs, the sample size for a learner's category depends on

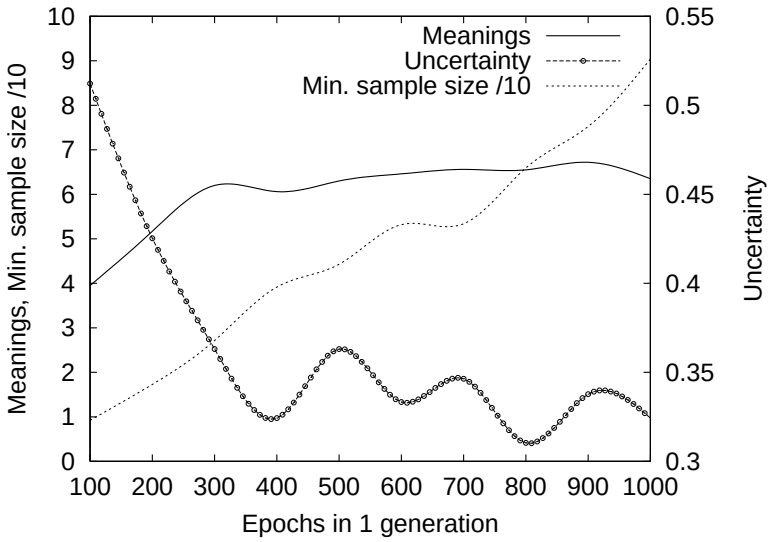

Figure 11.3: The influence of the number of learning epochs per generation on the stability of meanings in iterated intergenerational meaning transmission. Data for each number of learning epochs are averaged over 10 simulation runs with different random seeds. The measure *meanings* is the total number of the learner's meanings, averaged over all learning epochs in all generations. *Min. sample size/10* is the size of the sample size of the learner's criterion with the smallest sample set, averaged over all learning epochs in all generations (and scaled by 10). The *uncertainty* expresses the referential ambiguity of the teacher's descriptions (cf. Section 10.3), averaged in the same way as the two previous measures. The results show that smaller number of learning epochs causes smaller sample sets, which can lead to instability: some of the concepts get overspecialized and disappear (the number of total meanings gets smaller), others get overgeneralized (the uncertainty rises).

the probability of occurrence of instances of the teacher's category on the scene. This creates an implicit meaning bottleneck. In iterated learning models of grammar emergence, the learning bottleneck leads to the emergence of compositionality, because compositional rules are more likely to be transmitted through the bottleneck (Kirby and Hurford, 2001; Vogt, 2005). In our model, instances of more general categories are more likely to appear on the scene within the learning period than those of very specific categories or even categories representing individual objects. Also, our model shows the same frequency effects as those reported by Kirby and Hurford (2001): if examples of a very specific category appear often on the scene because of a biased random generator, they can get preserved over generations, otherwise they would probably die out.

Catching and amplifying randomly occurred regularities is the inherent property of iterated learning. While this property is desirable for the emergence of grammar, it can lead to distortion of meanings in our model. The smaller the sample, the bigger the chance that it will contain random correlations that are not a part of the original meaning and a covariance-based detector would not reconstruct the original meaning properly. Once a random correlation becomes a part of the meaning, it gets reinforced in the next generation, because the teacher will pick up as instances of the category only examples containing the correlation. This way the meaning gets overspecialized (see Figure 11.2a). Overspecialization is dumped by SVD-filtering (see Section 8.2.2) that captures the properties most invariant in the sample and filters out the others. However, a random invariance in a small sample can lead to overgeneralization due to truncating some relevant attributes (see Figure 11.2b).

Meaning transmission can be viewed as an evolutionary process with meanings as competing replicators. The selection pressure is imposed by the meaning bottleneck. Meanings pass through the bottleneck, if they are relevant to the environment. Special meanings describing situations that occur very rarely have smaller survival chances than frequently applicable general meanings. However, the model used in Experiment 1 corresponds with the replicator view only partially: meanings can die out, if they are no longer relevant, but there is no mechanism of creation of new meanings in the model. If the agents were suddenly relocated to a completely different environment, the teacher would remain silent because of the lack of adequate meanings, and the whole language would die out. To make the model more realistic, the teacher was allowed to coin new meanings and words in such situations. The experiment is described in the following section.

11.3 Experiment 2

The model setting was similar to that of Experiment 1, in that the teacher had to describe a randomly generated and modified scene to the learner. The teacher of the first generation started with the criteria for *triangle*, *square*, *small*, *big*, and *slim* (see Table 10.1). If, for some object on the scene, there was no criterion with an above-threshold activity, the object was approximately described by the word associated with a criterion returning the highest non-zero activity. If there was no criterion returning non-zero activity for the object, a new criterion (named by a new random word) was created with the object as the first example.

The learning process was iterated for 50 generations with 200, 500, or 1000 learning epochs in one generation. The results we got confirmed our replicator hypothesis. The meaning of *triangle* was stable and remained the same throughout all generations in all three versions of the experiment. The meaning of *square* remained stable in the experiments with 500 and 1000 learning epochs and died out in some simulation runs of the experiment with 200 learning epochs. In comparison to *triangle* and *square*, the criteria for *small*, *big*, and *slim* were more likely to return a non-zero activity for a random object. Hence, they were more often used in an approximate sense, which caused their instability.

Using criteria in approximate senses caused the extension of their receptive fields in the next generation. Meanings with under-threshold activity competed for selection and the meaning with highest activity was selected and extended subsequently. This created *rich-get-richer* dynamics (a positive feedback loop) and led to the formation of very general criteria. Indeed, in every version of the experiment, we observed the appearance and survival of general criteria[1] with meanings such as "objects with 2-5 vertices" or "objects with horizontal position *posX* between 0 and 50" applicable to all referents in all situations. Other newly-created meanings defeated in the competition had very small sample sets and have not survived.

[1] These included overgeneralized meanings of *small*, *big*, and *slim* as well as some newly created criteria.

Chapter 12

Discussion

12.1 Cognitive Plausibility and Implications

We have proposed a new semantic representation that combines advantages of symbolic, conceptual and subsymbolic levels of description (Gärdenfors, 1997). Locally tuned detectors that are the building blocks of our semantics can be constructed from scratch by utilizing statistical properties of their sample sets. As such, they are not far away from connectionist systems and could be implemented in this way (see comparison to Radial Basis Function Networks in the next section). Moreover, they have a high neural plausibility (Martin, 1991). Thanks to their geometric interpretation, locally tuned detectors provide a natural description of concepts with graded membership, fuzzy boundaries, prototype effects, similarity-based distances and potential for hierarchical relations and metaphoric mappings. Most importantly, they facilitate an ex-post analysis of the actual meaning of constructed representations.

We have also proposed, implemented and analyzed several original models of meaning construction. In the model of individual meaning construction, we have shown how ecologically relevant categories can be constructed from sensorimotor interactions with the environment. The resulting categories were relevant in that they reflected the structure and dynamics of the environment. The adequateness of the constructed representation was proved by the agent's ability to use it for predicting the results of its actions on objects. The categories of objects were organized by common interaction programs, in accordance with empirical findings of Rosch (1978) about basic-level categories. All action categories associated with some object category represented *affordances* of the object, i.e. the perceivable possibilities for acting on the object (Gibson, 1979). On the other hand, object and change

categories associated with action categories can be viewed as verb islands in line with the *verb island hypothesis* (Tomasello, 1992) stating that the first more complex lexical constructions of children are organized in verb-centered structures with verb-specific arguments.

Children learn by interacting with the world, but, at the same time, they are exposed to the linguistic production of their caregivers. Naming influences the children's conceptual organization and supports discovery of novel concepts (Waxman and Braun, 2005). We have explored this issue in our second computational model of two agents observing their dynamically changing environment. We have shown how a linguistic instruction from one agent (the teacher) accompanied by a non-verbal reference can lead to cross-situational construction of meanings by the other agent (the learner). The learner constructed different types of concepts represented by identification criteria, which were similar enough to those of the teacher, to be used for pragmatic purposes. A high similarity was achieved very rapidly, which was in line with the observed phenomenon of *fast mapping* (Carey and Bartlett, 1978).

However, meanings did not stay intact, when we let the acquisition process iterate in the third computational model. The causes of this instability were discussed in detail in Section 11.2.1. While a high similarity between teacher's and learner's meanings was maintained within each generation, meanings did change throughout the generations. These results suggest how real languages can change historically, while still preserving their communicative function.

Meanings constituted by simple structural relations and invariant attribute values happened to be more stable in the iterated transmission than meanings based on interval values of uncorrelated attributes. In our model, we used adjectives *big* and *small* with the meaning of having the size bigger or smaller than a fixed value. This is not realistic: most adjectives are semantically dependent on the nouns they modify (Warren, 1988), e.g. an adjective *big* refers to very different absolute measures in the phrases "big mouse" and "big elephant". The meanings of such adjectives are constituted by structural relations that are mapped onto a particular domain generated by the modified noun. We can speculate that the semantic dependency of adjectives, observed in real languages, is the result of the dynamics of the selection process within the iterated language transmission, where meanings based on structural relations are much more persistent.

If we allow extension of meanings to novel referents, the dynamics of the iterated transmission inevitably leads to the erosion toward more and more general meanings. This phenomenon was also observed in other models (Smith, 2001, 2005a). The occurrence of the meaning drift in our model

can be explained by the lack of other selection forces, as the meanings were only optimized for their expressive coverage. In real situations, utterances and their meanings serve pragmatic purposes including identification and discrimination. Hence, optimal meanings should reflect the trade-off between coverage and distinctiveness (Rosch et al., 1976).

To explore the nature of meaning creation mechanisms, we have deliberately studied each of them in isolation. However, in real situations, meaning formation processes are coupled and interact with each other, albeit they may operate on different timescales (Takáč, 2003a,b).

12.2 Related Works

There are lot of models related to our work in various aspects. The similarities and differences can be analyzed on the level of representation, learning mechanisms and the overall dynamics.

12.2.1 Representation

In the models of language bootstrapping, meanings are usually represented by collections of intervals from discrimination trees (Steels, 2000; Smith, 2005a), prototypes in one-dimensional (Vogt and Divina, 2007) or multidimensional (Vogt, 2005) conceptual spaces (Gärdenfors, 2000), adaptive networks (Steels and Belpaeme, 2005), or weight configurations in artificial neural networks (Borghi et al., 2005; Cangelosi, 2005).

Nodes of a *discrimination tree* represent features – subintervals of the range of a particular sensory channel. The initial range $[0, 1]$ is adaptively refined, based on the results of discrimination games. In some models, if a single feature cannot identify a topic referent uniquely, a set of features is chosen. However, the discrimination trees are mutually independent and the construction of the feature set is situational and not persistent. A discrimination tree related to a particular sensory channel can be likened to a locally tuned detector only based on one attribute in our model. However, locally tuned detectors can also persistently represent meanings based on multiple attributes and their correlations, therefore our representation has bigger expressive power than discrimination trees.

Prototypes in one-dimensional spaces related to perceptual features have the same expressive power as features in discrimination trees. Prototypes in a multidimensional conceptual space have the potential to capture correlations of attributes. However, in the above-mentioned models, categories in a multidimensional space are constructed by placing prototypes on each di-

mension separately and combining them together (Vogt, 2005), which makes this representation insensitive to unequal importance of dimensions for a category. The density of each dimension is the same for all categories, because a category of an object is constructed as a vector of the closest prototypes on each dimension and the prototypes generate a grid in the conceptual space. Categories based on correlated attributes could in principle be represented as multiple nodes of the grid, but this is not used in the reviewed models. Hence, the advantage of our representation is sensitivity to the importance of each dimension for the category membership and a power to persistently represent multidimensional concepts based on inter-correlations of attributes.

Adaptive networks used by Belpaeme (2002); Steels and Belpaeme (2005) are most closely related to our representation. An adaptive network consists of a set of locally reactive units, each with a Gaussian activation function centered at some point (widths of all Gaussians are the same and fixed to some experimentally determined constant). The resulting activity of the network is a weighted linear combination of activities of all reactive units applied to a common input. Each category is represented by one adaptive network; an input is categorized as a member of the category represented by the network giving the highest activity.

Adaptive networks can be considered variants of Radial Basis Function Networks (RBFN, Poggio and Girossi, 1990b).[1] It has been proved that RBFN is a universal approximator in that it can approximate any multivariate continuous function, given a sufficient number of reactive units (Poggio and Girossi, 1990a). A single reactive unit of RBFN cannot capture correlations of attributes, but the whole network can. A hyperelliptic receptive field of a category represented by one locally tuned detector based on Mahalanobis metric in our approach can be covered by many locally reactive units of RBFN with suitably chosen positions of their centers.[2] The advantage of a multi-unit RBFN over our single-unit locally tuned detectors is that RBFN can also represent categories whose receptive fields are not convex or even consist of disconnected regions. However, such categories are not considered natural (Gärdenfors, 2000). Representing one natural category by multiple units (corresponding to multiple prototypes) is less economic and does not

[1] Belpaeme (2002, p. 57) have chosen the name *adaptive network* instead of radial basis function networks, to emphasize the difference between RBFN, which are trained to fit a function using a learning method, and adaptive networks, which are adapted according to their performance in discrimination games. See the next section for details.

[2] Each locally reactive unit of RBFN covers a hypersphere, i.e. it corresponds to a locally tuned detector with the common variance based Euclidean metric $d_{L_2,\sigma}$ (see Section 8.2.1). Generalized RBFN (Poggio and Girossi, 1990b) use units with covariances, which are equivalent to the pseudoinverse version of our covariance-based locally tuned detectors.

have to be cognitively plausible.

However, having multi-detector representation of categories could be a good compromise between exemplar-based (Nosofsky, 1984) and prototype-based (Rosch, 1978) approaches to categorization. During the category induction, sufficiently similar examples would update the same detector, while a new detector could be added for more distant examples. This approach would support multiple levels of abstraction, where the multi-unit criterion represents a superordinate category consisting of several basic-level subcategories (see Section 2.4.1) and could also cope with synonymy.

Last but not least, the representation proposed in this thesis is in accordance with the ideas of Harnad (1990): our perceptual frames are iconic projections of perceived objects, and elementary identification criteria correspond to categorical representations that pick up invariant features of categories. Although we encode frames and detectors in symbolic fashion, they both clearly have nonsymbolic correlates. Words associated with the criteria in the language level are Harnad's elementary symbols. Hierarchical identification criteria and propositional associations of criteria representing action knowledge correspond to Harnad's higher order symbolic representations.

12.2.2 Learning Mechanisms

The crucial difference between RBFN, adaptive networks and our locally tuned detectors is in the way they learn.

Several mechanisms of learning have been proposed for RBFN (Haykin, 1999; Beňušková, 2002b). While the hidden layer's activation functions evolve slowly in accordance with some non-linear strategy, the output layer's weights adjust themselves rapidly through a linear optimization strategy (Haykin, 1999, p. 298). The number of locally reactive units is decided experimentally beforehand. Location of centers of the units can be determined either by random selection from the sample set, or in a self-organized way by *k-means clustering algorithm*,[3] or by a supervised gradient-descent error-minimizing procedure. The width of Gaussians is chosen as a multiple of average distances between the centers to allow for a small overlap between radial functions, or it can be determined by the supervised gradient-descent procedure. The linear output-layer weights are determined by a supervised error-minimizing procedure.

Wettschereck and Dietterich (1992) have compared the performance of RBFN with fixed centers to that of generalized RBFN with adjustable centers determined by supervised learning on NETtalk task focusing on mapping En-

[3]For details, see e.g. Beňušková (2002b).

glish spelling into its phonetic pronounciation. The NETtalk experiment was originally carried out by Sejnowski and Rosenberg (1987) using multi-layered perceptron trained with the back-propagation algorithm. The results of experimental comparison of the generalization performance have shown that generalized RBFN with computationally more intensive supervised learning of the parameters of locally reactive units as well as the output-layer weights performed substantially better than the original multi-layered perceptrons, while RBFN with self-organized locations of centers of locally reactive units and supervised learning of the output-layer weights did not achieve the performance level of the original perceptrons (Haykin, 1999, p. 325).

The adaptive networks of Steels and Belpaeme (2005) are not trained in the same way as RBFN. Rather, they are adapted by adding or removing a locally reactive unit and by changing weights of the units (the width and centers of the units remain unchanged). When an adaptive network that represents a category is created, it consists of a single locally reactive unit centered on the first example. The adaptation of networks is guided by their success or failure in discrimination games[4] (see Section 4.4.1). If no discriminating category could be found for some topic, either a new category (network with a single unit) is created, or the best matching category is adapted by adding a new locally reactive unit centered on the topic. If a discrimination game is successful, the weight of each locally reactive unit of the discriminating category is increased proportionally to the unit's activation. After every discrimination game, the weights of all the locally tuned units of all categories of an agent are decreased with a non-negative decay. When the weight of some locally tuned unit is lower than a certain threshold, the locally tuned unit is removed from the adaptive network. When no more locally tuned units are associated with the network, the whole network is removed. Hence, adaptive networks use exemplar-based representation of categories with a forgetting mechanism based on stimulation frequencies.

The crucial difference between training a RBFN and adapting an adaptive network is that the latter is incremental (it does not need to have the whole sample set in advance) and faster, having a reasonable level of performance even after seeing first few examples.

Learning in our models corresponds to finding the values of parameters (the center and the covariance matrix) of a single unit generalized RBFN. However, it is not so computationally costly as non-linear optimization of RBFN, because it is instance-based and incremental. It is based on extracting cross-situational similarities between examples of a category. Input frames

[4]Other models of Steels' group (for overview see Steels, 2000) are based on the same principle of learning from feedback about success or failure in various games.

are considered examples of the same category, if they can be interacted with in the same way (in the case of individual learning), or if they were named by the same word (in case of social learning).

A seminal model of cross-situational learning was published by Siskind (1996). In this model, the algorithm learns mappings between word symbols and conceptual symbols such as GO, **John**, **ball**. A hypothesis set of all possible conceptual symbols is given in advance, and "lexical acquisition is simply a process of learning the mapping between two pre-existing mental representation languages" (Siskind, 1996, p. 47).

In this aspect, we can view our model as cross-situational learning of meanings of Siskind's atomic conceptual symbols, while learning in the model of Siskind works more on the sentence level by eliminating meaning mappings incoherent across situations. Siskind's model deals with the referential indeterminacy, noise and homonymy by employing the mutual exclusivity assumption (Markman, 1992).

Cross-situational learning is also used in the model of Smith (2005a). Semantic hypotheses (represented by nodes of discrimination trees) are not given in advance, but are constructed by playing discrimination games prior to the language acquisition phase. In the language acquisition phase, agents learn mappings between the constructed meanings and words from word-meaning co-occurrence frequencies. Cross-situational learning is combined with learning based on corrective feedback in the model of Divina and Vogt (2006); Vogt and Divina (2007). Mathematical properties of cross-situational learning are analyzed by Smith et al. (2006).

Among recent connectionist models of action-based categorization, that of Borghi et al. (2005) seems to be most closely related to our model of individual interactionist meaning construction. In their model, an organism with a visual system and a two-segment arm (simulated by a neural network) reaches different points in space, depending on the object seen and on the context. Constructed categories reflect characteristics of the output actions to be performed rather than perceptual characteristics of the input. However, the organism is selected from a population of non-learning neural networks by genetic algorithm, which is not plausible as a model of ontogenetic acquisition of categories.

The role of social learning in the acquisition of concepts and language has been studied by Steels and Kaplan (2001a). In their experiment, a human teacher (mediator) interacted with a Sony AIBO robot, trying to teach it the names of three objects. The perceptual input of the robot was in the form of camera images taken from different angles and under different

light conditions. The robot used simple instance-based learning[5] and this was compared with unsupervised clustering techniques. The experiment has shown that categories obtained by unsupervised techniques did not match the three objects, while categorization directed by naming has been much more successful. Active and rich social interactions served the role of narrowing the context and reducing the noise. If the teacher just estimated at which object the robot was looking, and uttered a name for it, the input was much more noisy and the performance deteriorated.

The experiment have led the authors to the conclusion similar to ours: naming has a beneficial effect on the categorization process. In comparison to our approach, they used much simpler learning algorithm that was not sensitive to unequal importance of dimensions. The authors admit that if there were more objects on the scene represented by perceptual images in more dimensional space, methods for computing correlations of dimensions with categories and the intercorrelations among dimensions should be employed (Steels and Kaplan, 2001a, p. 26). Our algorithm is sensitive to multidimensional intercorrelations.

12.2.3 Iterated Intergenerational Transmission

In the paradigmatic iterated learning model (Kirby and Hurford, 2001) focused on the emergence of compositional structures on the syntax level, meanings were pre-defined and artificial structures. This has been refined in the iterated learning model of Vogt (2005), in which meanings were created in discrimination games. Cross-situational learning of meanings was combined with iterated vertical transmission in the model of Smith (2005a). Although, in this model, meanings were created individually by each agent in discrimination games, the experiment led to results similar to ours (high intra-generational meaning similarity, decreasing inter-generational meaning persistence, and the drift toward more general meanings).

12.3 Limits and Future Work

Our models have been simplified in many aspects. First, categories were constructed by taking into account attributes common to *all* examples. This approach works well for basic level categories, but can be problematic for some superordinate categories (remember that basic-level is the most general

[5]Previous instances of category members were stored in the form of histograms. Classification took place by a nearest neighbor algorithm evaluating the distance of two histograms by a χ^2-divergence measure.

level, at which a common perceptual image and a common motor program can be created for members of a category). Moreover, instances of some concepts are related by *family resemblances* rather than by common properties of all members, as exemplified by Wittgenstein (1953) with the concept of a *game*. In its current form, the model cannot cope with homonyms and with noise. If a sample set erroneously contained an instance with a set of attributes completely different from other instances, it would result in a category with an empty attribute set. This could be amended by recording frequencies of attributes and by splitting concepts in case of homonymy detection.

Second, our models do not account for hierarchic and taxonomic relations that exist among real concepts.[6] Also, the semantics of verbs in our model is based on criteria of one-step changes of attribute values. Representation of larger sequences of changes may be necessary for some verbs. For other verbs, discrete sequences may be insufficient at all and some kind of continuous representation of the dynamics (e.g. phase portraits, see Section 4.5.2) may be required.

Third, in the model of social learning of concepts, an explicit reference (focus) to instances of the named category was given along with the linguistic input.[7] We used this simplification deliberately, in order to show that even in the absence of *referent* indeterminacy, the learner has to solve *sense* indeterminacy, because different words can describe different aspects of the same (known) referent. However, in the later phases of the language acquisition process, the explicit reference could be substituted by the inference from the linguistic or pragmatic context (e.g. the utterance "*big X*" can narrow the context to big objects on the scene).

Meaningful categories should be useful for the agent in achieving its goals (Nehaniv, 2000). In our action-based model, the agent had no goals and performed actions randomly. The next research step is to endow the agent with needs, need-driven goals and an action planning mechanism.[8] Also, we plan to study the interplay of the meaning construction mechanisms in a model with coupled individual and social learning.

According to the syntactic bootstrapping theory (Lidz et al., 2004), children acquiring a language use grammatical cues to constrain possible meanings of words. The principles we used for acquisition of meanings of single

[6]Possibilities of hierarchic relations between identification criteria were explored by Višňovská (2007).

[7]In the first, bootstrapping, phase of children's language acquisition, the focus (non-verbal reference) is established by joint attention of the child and the mother, gaze following and pointing (Tomasello and Farrar, 1986).

[8]Preliminary attempts heading in this direction can be found in the work of Jankovič (2007).

words could also be applied to multi-word noun phrases: each word of a phrase is assumed to denote a different aspect of the referent. In the acquisition model with no grammar, the word order of the phrase is unimportant and induction from the phrase "*left big triangle*" (or any of its permutations) has the same effect as three subsequent inductions from single words. Also, even if the agent has acquired the correct meanings of words such as "*cat*", "*on*", "*hot*", "*tin*", and "*roof*" in the single-word induction setting, it cannot understand the meaning of a phrase "*cat on hot tin roof*", unless it knows the rules of grammatical word composition. Incorporating some form of grammar into our model is a topic for future research.

Chapter 13

Conclusion

The principal goals of this thesis were: (1) to formulate a theory of interaction-based meaning construction, (2) propose a formal representation of various types of meanings, and (3) study mechanisms of individual and social construction of the proposed representation by computational modeling methodology. All these goals were fulfilled.

We described a grounded cognitive semantics for representing concepts of objects, properties, relations, changes, complex situations and events, based on *identification criteria*. The identification criteria are constructed individually by each agent, based on interactions with the environment and other agents. Unlike in most of the related models, construction of criteria is based on cross-situational *similarities* among instances of a category rather than on *differences* between a chosen object and other objects present on the scene of communication. We argue that categories constructed for the purpose of identification rather than discrimination are more suitable for the detached use of language (talking about things not present here and now).

Learning in our models is incremental and permanent. The learning mechanism is sensitive to correlations of attributes of instances with categories and the intercorrelations among attributes. We have implemented and experimentally tested meaning construction by individual and social learning, and explored the dynamics of meanings in iterated intergenerational transmission.

We would like to emphasize that, in the presented models, categories are not given and interpreted by an external designer, but are constructed by and meaningful to the agents themselves. Such models can have important practical applications in the areas involving agents that need to coordinate their activities in unknown, dynamic and open environments. As all possible meanings cannot be anticipated in design-time, the agents' ability to acquire (and continuously reconstruct) relevant meanings is critical.

Resumé

V tejto dizertačnej práci sme sformulovali teóriu významov založenú na interakciách s prostredím tak, aby bola aplikovateľná nielen na ľudskú jazykovú komunikáciu, ale aj na interakcie predverbálnych živých organizmov i umelých systémov. Originálnym prínosom práce je návrh sémantickej reprezentácie založenej na prirodzených podobnostiach tak, že jednotný formalizmus tzv. identifikačných kritérií umožňuje konštrukciu významov reprezentujúcich nielen statické objekty, ale aj ich vlastnosti, vzťahy, dynamické zmeny, situácie a udalosti. Treba zdôrazniť, že významy nie sú dané vopred ani nie sú interpretované externým pozorovateľom, ale sú konštruované (naučené) samotnými agentmi, pre ktoré majú inherentný význam. Konštrukcia významov je inkrementálna a permanentná. Nami navrhnuté mechanizmy konštrukcie významov založené na senzomotorických a sociálnych interakciách sme implementovali a experimentálne overili.

Prvá séria experimentov preverila funkčnosť navrhnutého mechanizmu vytvárania kategórií na základe senzomotorickej interakcie s prostredím. Vytvorené kategórie umožnili efektívnu predikciu dôsledkov agentových akcií. Navyše použitá reprezentácia sa ukázala pre tento účel vhodnejšou ako prototypy, vďaka citlivosti lokálnych detektorov na rôznu dôležitosť atribútov pre príslušnosť ku kategórii.

V druhej sérii experimentov sme ukázali, ako si učiaci sa agent na základe verbálnej inštrukcie spolu s neverbálnou referenciou skonštruuje významy, ktoré sú pre plnenie pragmatických cieľov dostatočne podobné významom učiteľa. Vysoká podobnosť bola dosiahnutá rýchlo, čo je v súlade s pozorovaným fenoménom rýchleho učenia (fast mapping) u detí.

Tretia séria experimentov bola zameraná na skúmanie stability navrhnutej reprezentácie v medzigeneračnom prenose. Ukázalo sa, že medzigeneračný prenos významov môžeme chápať ako evolučný proces, v ktorom sú významy replikátormi súťažiacimi o prežitie, pričom selekčným tlakom je zúžený profil prenosu významov. Významy prejdú cez zúžený profil, ak sú relevantné prostrediu (teda ich inštancie sa v ňom vyskytujú dostatočne často). Ak necháme proces akvizície významov iterovať, významy neostanú nezmenené.

Aj keď v každej generácii ostane vysoká podobnosť medzi význammi učiteľa a žiaka, medzigeneračne sa významy budú posúvať a vyvíjať. Tieto výsledky zodpovedajú tomu, že jazyky podliehajú historickým premenám bez toho, aby stratili svoju dorozumievaciu funkciu.

Výhodou navrhnutého učiaceho mechanizmu oproti iným modelom je jeho citlivosť nielen na korelácie atribútov inštancií s príslušnosťou ku kategórií aj na vzájomné korelácie medzi atribútmi. Na rozdiel od mnohých existujúcich prístupov, konštrukcia identifikačných kritérií je založená na medzisituačných podobnostiach inštancií konceptov, a nie na rozdieloch medzi zvoleným objektom a ostatnými aktuálne prítomnými objektmi, čo je predpoklad pre situačne nezávislé používanie jazyka.

Bibliography

Akhtar, N., Montague, L., 1999. Early lexical acquisition: the role of cross-situational learning. First Language 19, 347–358.

Arrowsmith, D. K., Place, C. M., 1990. An Introduction to Dynamical Systems. Cambridge University Press, Cambridge.

Austin, J. L., 1962. How to Do Things With Words. Harvard University Press, Cambridge, MA.

Bailey, D., Feldman, J., Narayanan, S., Lakoff, G., 1997. Modeling embodied lexical development. In: Proceedings of the 19th Cognitive Science Society Conference. pp. 19–24.

Bailey, D. R., 1997. When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs. Ph.D. thesis, University of California at Berkeley.

Baldwin, J. M., 1896. A new factor in evolution. American Naturalist 30, 441–451.

Balkenius, C., 1999. Are there dimensions in the brain? In: Spinning Ideas, Electronic Essays Dedicated to Peter Gärdenfors on His Fiftieth Birthday. Retrieved from `http://www.lucs.lu.se/spinning/categories/cognitive/Balkenius/Balkenius.pdf`.

Barsalou, L. W., 1999. Perceptual symbols systems. Behavioral and Brain Sciences 22, 577–660.

Bartmiński, J., Tokarski, R., 1986. Językowy obraz świata a spójność tekstu. In: Dobrzyńska, T. (Ed.), Teoria tekstu. Zbiór studiów. Wrocław, pp. 65–81.

Batali, J., 1998. Computational simulations of the emergence of grammar. In: Hurford, J. R., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches

to the Evolution of Language: Social and Cognitive Bases. Cambridge University Press, Cambridge, pp. 405–426.

Bealer, G., 1989. On the identification of properties and propositional functions. Linguistics and Philosophy 12 (1), 1–14.

Belpaeme, T., 2002. Factors influencing the origins of colour categories. Ph.D. thesis, Vrije Universiteit Brussel, Artificial Intelligence Lab.

Beňušková, L., 2002a. Kognitívna neuroveda [Cognitive neuroscience]. In: Rybár, J., Beňušková, L., Kvasnička, V. (Eds.), Kognitívne vedy. Kalligram, Bratislava, pp. 47–104.

Beňušková, L., 2002b. Umelé neurónové siete [Artificial neural networks]. In: Návrat, P., et al., Umelá inteligencia, STU Publishing, Bratislava, pp. 161–189.

Beňušková, L., 2005. Kde sa jazyk stretáva s vedomím [Where language meets consciousness]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 235–261.

Beňušková, L., Kasabov, N., 2007. Computational Neurogenetic Modeling. Springer, New York.

Bergen, B., Chang, N., 2003. Embodied construction grammar in simulation-based language understanding. In: Ostman, J. O., Fried, M. (Eds.), Construction Grammar(s): Cognitive and Cross-Language Dimensions. Johns Benjamins, pp. 147–190.

Berlin, B., Kay, P., 1969. Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley, CA.

Binswanger, L., 1942/1993. Grundformen und Erkenntnis menschlichen Daseins, Vol. 2 of selected works (Ausgewählte Werke, ed. Max Herzog and Hans-Jürg Braun). Roland Asanger Verlag, Heidelberg.

Bloom, P., 2000. How Children Learn the Meanings of Words. MIT Press, Cambridge, MA.

Bodík, P., Takáč, M., 2003. Formation of a common spatial lexicon and its change in a community of moving agents. In: Tessem, B., Ala-Siuru, P., Doherty, P., Mayoh, B. (Eds.), Frontiers in AI: Proceedings of the Eighth Scandinavian Conference on Artificial Intelligence (SCAI'03). IOS Press, Amsterdam, pp. 37–46.

Booth, A. E., Waxman, S. R., 2002. Object names and object functions serve as cues to categories for infants. Developmental Psychology 38 (6), 948–957.

Borghi, A. M., Parisi, D., Di Ferdinando, A., 2005. Action and hierarchical levels of categories: A connectionist perspective. Cognitive Systems Research 6, 99–110.

Bratman, M., 1987. Intention, Plans, and Practical Reason. Harvard University Press, Cambridge, MA.

Briscoe, T. (Ed.), 2001. Linguistic Evolution Through Language Acquisition: Formal and Computational Models. Cambridge University Press, Cambridge, U. K.

Brooks, R. A., 1990. Elephants don't play chess. Robotics and Autonomous Systems 6 (1–2), 3–15.

Brooks, R. A., 1991a. Intelligence without reason. In: Proceedings of the Twelfth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 569–595.

Brooks, R. A., 1991b. Intelligence without representation. Artif. Intell. 47 (1-3), 139–159.

Brooks, R. A., 1991c. The role of learning in autonomous robots. In: COLT '91: Proceedings of the fourth annual workshop on Computational learning theory. Morgan Kaufmann, San Francisco, CA, pp. 5–10.

Cangelosi, A., 1999a. Evolution of communication using combination of grounded symbols in populations of neural networks. In: Proceedings of IJCNN99 International Joint Conference on Neural Networks. Vol. 6. IEEE Press, Washington, DC, pp. 4365–4368.

Cangelosi, A., 1999b. Modeling the evolution of communication: From stimulus associations to grounded symbolic associations. In: Floreano, D., Nicoud, J., Mondada, F. (Eds.), Proceedings of the 5th European Conference on Advances in Artificial Life. Springer-Verlag, Berlin, pp. 654–663.

Cangelosi, A., 2005. Approaches to grounding symbols in perceptual and sensorimotor categories. In: Cohen, H., Lefebvre, C. (Eds.), Handbook of Categorization in Cognitive Science. Elsevier, pp. 719–737.

Cangelosi, A., 2006. The grounding and sharing of symbols. Pragmatics and Cognition 14 (2), 275–285.

Cangelosi, A., Parisi, D., 2001. How nouns and verbs differentially affect the behavior of artificial organisms. In: Moore, J. D., Stenning, K. (Eds.), Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, London, pp. 170–175.

Cangelosi, A., Parisi, D., 2004. The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. Brain and Language 89 (2), 401–408.

Carey, S., Bartlett, E., 1978. Acquiring a single new word. Papers and Reports on Child Language Development 15, 17–29.

Carpenter, R., 2007. Jabberwacky – live chatbot, available online at `http://www.jabberwacky.com`.

Čerňanský, M., Makula, M., Beňušková, L., 2007. Organization of the state space of a simple recurrent neural network before and after training on recursive linguistic structures. Neural Networks 20 (2), 236–244.

Chandler, D., 2007. Semiotics: the basics, 2nd Edition. Routledge, London, New York.

Chang, N., 2004. Constructing grammar: A computational model of the emergence of early constructions. Ph.D. thesis, University of California at Berkeley.

Chierchia, G., 1999. Linguistics and language. In: Wilson, R. A., Keil, F. C. (Eds.), The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge, MA, pp. xci–cix.

Chomsky, N., 1957. Syntactic Structures. Mouton and Co, The Hague.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., 1980. Rules and Representations. Columbia University Press, New York.

Chomsky, N., 1986. Knowledge of Language. Its Nature, Origin, and Use. Convergence. Praeger, New York/Westport/London.

Clark, E., 1987. The principle of contrast: A constraint on language acquisition. In: MacWhinney, B. (Ed.), Mechanisms of language acquisition. Lawrence Erlbaum Assoc., Hillsdale, NJ, pp. 1–33.

Cohen, P., 1998. Dynamic maps as representations of verbs. In: Proceedings of the 13th Biennial European Conference on Artificial Intelligence. pp. 145–149.

Cohen, P., Oates, T., Atkin, M., Beal, C., 1996. Building a baby. In: Cottrell, G. W. (Ed.), Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 518–522.

Cole, D., 2004. The chinese room argument. In: Zalta, E. N. (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2004 Edition. `http://plato.stanford.edu/archives/fall2004/entries/chinese-room`.

Coles, L. S., 1969. Talking with a robot in English. In: Donald, E., Lewis, M. N. (Eds.), Proceedings of the International Joint Conference on Artificial Intelligence. Walker, Washington, D.C., pp. 587–596.

Csontó, J., 2001. Umelý život [Artificial life]. In: Mařík, V., et al. (Eds.), Umělá inteligence. Vol. 3. Academia, Praha, pp. 76–116.

Csontó, J., Palko, M., 2002. Umelý život [Artificial Life]. ELFA, Košice.

Curtiss, S. (Ed.), 1977. Genie: Psycholinguistic Study of a Modern-day "Wild Child". Academic Press Inc., London.

Dawkins, R., 1976. The Selfish Gene, 2nd Edition. Oxford University Press, Oxford.

de Chardin, P. T., 1956. Le phénomène humain. Les Éditions du Seuil, Paris.

de Jong, E., Vogt, P., 1998. How should a robot discriminate between objects? A comparison between two methods. In: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior SAB'98. MIT Press, Cambridge, MA, pp. 86–91.

de Jong, E. D., 1999. Autonomous concept formation. In: Dean, T. (Ed.), IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, pp. 344–349.

de Jong, E. D., 2000. Autonomous formation of concepts and communication. Ph.D. thesis, Vrije Universiteit Brussel.

de Saussure, F., 1916/1974. Course In General Linguistics. Fontana/Collins, London.

Deacon, T. W., 1997. The Symbolic Species: The Co-evolution of Language and the Brain. W.W. Norton & Co., New York, N.Y.

Deely, J., 2001. Umwelt. Semiotica 134 (1/4), 125–135.

Divina, F., Vogt, P., 2006. A hybrid model for learning word-meaning mappings. In: Vogt, P., Sugita, Y., Tuci, E., Nehaniv, C. (Eds.), Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication. Springer Verlag, Berlin/Heidelberg, pp. 1–15.

Dore, J., 1975. Holophrases, speech acts and language universals. Journal of Child Language 2, 21–40.

Everitt, B. S., Landau, S., Leese, M., 2001. Cluster Analysis. Arnold, London.

Farkaš, I., 2003. Lexical acquisition and developing semantic map. Neural Network World 13 (3), 235–245.

Farkaš, I., 2005. Konekcionistické modelovanie jazyka [Connectionist language modelling]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 262–305.

Farkaš, I., Li, P., 2001. A self-organizing neural network model of the acquisition of word meaning. In: Altmann, E., Cleeremans, A., Schunn, C., Gray, W. (Eds.), Proceedings of the 4th International Conference on Cognitive Modeling. Fairfax, VA, pp. 67–72.

Fauconnier, G., 1985. Mental Spaces: Aspects of Meaning Construction in Natural Language. MIT Press, Cambridge, MA.

Feldman, J., 2006. From Molecule to Metaphor: A Neural Theory of Language. MIT Press, Cambridge, MA.

Fikes, R., Nilsson, N., 1971. STRIPS: A new approach to the application of theorem proving to problem solving. Tech. Rep. 43r, AI Center, SRI International, Menlo Park, CA.

Fillmore, C. J., 1982. Frame semantics. In: Linguistics in the Morning Calm. Hanshin Pub. Co., Seoul, pp. 111–137.

Fodor, J. A., 1975. The Language of Thought. Harvard University Press, Cambridge, MA.

Fodor, J. A., 1981. Representations: Philosophical Essays on the Foundations of Cognitive Science. MIT Press, Cambridge, MA.

Fodor, J. A., Pylyshyn, Z. W., 1988. Connectionism and cognitive architecture: a critical analysis. In: Connections and symbols. MIT Press, Cambridge, MA, pp. 3–71.

Frege, G., 1892/1952. Über Sinn und Bedeutung. Zeitschrift für Philosophie und Philosophische Kritik 100, 25–50, (Translated by M. Black under the title "On Sense and Reference", in P. Geach and M. Black, Translations from the Philosophical Writings of Gottlob Frege, Oxford, 1952.).

Gazzaniga, M. S. (Ed.), 1999. The New Cognitive Neurosciences, 2nd Edition. The MIT Press, Cambridge, MA.

Gelder, T. J., 1999. Dynamic approaches to cognition. In: Wilson, R. A., Keil, F. C. (Eds.), The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge, MA, pp. 244–246.

Gibbs, R. W., 2006. Embodiment and Cognitive Science. Cambridge University Press, Cambridge.

Gibson, J. J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin, Boston.

Gold, E. M., 1967. Language identification in the limit. Information and Control 10 (5), 447–474.

Goodman, N., 1955. Fact, Fiction, and Forecast. Harvard University Press, Cambridge, MA.

Goodwin, B. C., 1978. A cognitive view of biological process. Journal of Social and Biological Structures 1, 117–125.

Gärdenfors, P., 1996a. Cued and detached representations in animal cognition. Behavioral Processes 35, 263–273.

Gärdenfors, P., 1996b. Language and the evolution of cognition. In: Rialle, V., Fisette, D. (Eds.), Penser l'esprit: Des sciences de la cognition a‘ une philosophie cognitive. Presses Universitaires de Grenoble, Grenoble, pp. 151–172.

Gärdenfors, P., 1997. Symbolic, conceptual and subconceptual representations. In: Cantoni, V., di Gesù, V., Setti, A., Tegolo, D. (Eds.), Human and Machine Perception: Information Fusion. Plenum Press, New York, pp. 255–270.

Gärdenfors, P., 2000. Conceptual Spaces. MIT Press, Cambridge, MA.

Gärdenfors, P., 2004. Cooperation and the evolution of symbolic communication. In: Oller, K., Griebel, U. (Eds.), The Evolution of Communication Systems. MIT Press, Cambridge, MA, pp. 237–256.

Harm, M., 2002. Building large scale distributed semantic feature sets with WordNet. Tech. Rep. PDP-CNS-02-1, Carnegie Mellon University.

Harnad, S., 1990. The symbol grounding problem. Physica D 42, 335–346.

Harnad, S., 2005. Language and the game of life. Commentary on "Coordinating perceptually grounded categories through language. A case study for colour." L. Steels & T. Belpaeme. Behavioral and Brain Sciences 28 (4), 497–498.

Hassoun, M. H., 1995. Fundamentals of Artificial Neural Networks. MIT Press, Cambridge, MA.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, 2nd Edition. Prentice Hall, Upper Saddle River, NJ.

Hulth, N., Grenholm, P., 1998. A distributed clustering algorithm. Lund University Cognitive Studies 74.

Icogno, 2007. What AI techniques does Jabberwacky use?, retrieved from `http://www.icogno.com/what_ai_techniques.html`.

Ientilucci, E. J., 2003. Using the singular value decomposition, retrieved from `http://www.cis.rit.edu/~ejipci/Reports/svd.pdf`.

Jankovič, P., 2007. Sémantická reprezentácia pragmatických znalostí [Semantic representation of pragmatic knowledge]. Master's thesis, Comenius University, Bratislava.

Johnson, S. P., Amso, D., Slemmer, J. A., 2003. Development of object concepts in infancy: Evidence for early learning in an eye tracking paradigm. Proceedings of the National Academy of Sciences USA 100, 10568–10573.

Kay, P., Kempton, W., 1984. What is the Sapir-Whorf hypothesis? American Anthropologist 86 (1), 65–79.

Kelemen, J., 1994. Strojovia a agenty. Archa, Bratislava.

Kelemen, J., Ftáčnik, M., Kalaš, I., Mikulecký, P., 1992. Základy umelej inteligencie [Foundations of Artificial Inteligence]. Alfa, Bratislava.

Kirby, S., 2000. Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: Knight, C. (Ed.), The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form. Cambridge University Press, Cambridge, pp. 303–323.

Kirby, S., 2002. Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe, T. (Ed.), Linguistic Evolution through Language Acquisition: Formal and Computational Models. Cambridge University Press, Cambridge, Ch. 6, pp. 173–204.

Kirby, S., Hurford, J., 1997. Learning, culture and evolution in the origin of linguistic constraints. In: Husbands, P., Harvey, I. (Eds.), Proceedings of the Fourth European Conference on Artificial Life (ECAL97). MIT Press, Cambridge, MA, pp. 493–502.

Kirby, S., Hurford, J., 2001. The emergence of linguistic structure: an overview of the iterated learning model. In: Parisi, D., Cangelosi, A. (Eds.), Computational Approaches to the Evolution of Language and Communication. Springer-Verlag, Berlin, pp. 121–148.

Kováč, L., 1986. Úvod do kognitívnej biológie [Introduction to cognitive biology]. Biologické listy 51 (3), 172–190.

Kováč, L., 2000. Fundamental principles of cognitive biology. Evolution and Cognition 6, 51–69.

Kováč, L., 2003. Ľudské vedomie je produktom evolučnej eskalácie emocionálneho výberu [Human consciousness is a product of evolutionary escalation of emotional selection]. In: Kelemen, J. (Ed.), Kognice a umělý život III. Slezská univerzita, Opava, pp. 75–93.

Kováč, L., 2006. Princípy molekulárnej kognície [Principles of molecular cognition]. In: Kelemen, J., Kvasnička, V. (Eds.), Kognice a umělý život VI. Slezská univerzita, Opava, pp. 215–222.

Kripke, S. A., 1959. A completeness theorem in modal logic. Journal of Symbolic Logic 24, 1–15.

Kripke, S. A., 1963. Semantical considerations on modal logic. Acta Philosophica Fennica 16, 83–94.

Kuipers, B., 1994. Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge. MIT Press, Cambridge, MA.

Kuipers, B., Beeson, P., Modayil, J., Provost, J., 2006. Bootstrap learning of foundational representations. Connection Science 18 (2), 145–158.

Kvasnička, V., Beňušková, L., Pospíchal, J., Farkaš, I., Tiňo, P., Král, A., 1997. Úvod do teórie neurónových sietí [An Introduction into the Theory of Neural Networks]. IRIS, Bratislava.

Kvasnička, V., Pospíchal, J., 1999. An emergence of coordinated communication in populations of agents. Artificial Life 5 (4), 319–342.

Kvasnička, V., Pospíchal, J., 2002. Konekcionizmus a modelovanie kognitívnych procesov [Connectionism and modeling of cognitive processes]. In: Rybár, J., Beňušková, L., Kvasnička, V. (Eds.), Kognitívne vedy. Kalligram, Bratislava, pp. 257–345.

Kvasnička, V., Pospíchal, J., 2005. O nevyhnutnosti univerzálnej gramatiky [About the necessity of universal grammar]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 361–390.

Kvasnička, V., Pospíchal, J., Tiňo, P., 2000. Evolučné algoritmy [Evolutionary algorithms]. STU Publishing, Bratislava.

Lakoff, G., 1987. Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. University of Chicago Press, Chicago.

Lakoff, G., Johnson, M., 1980. Metaphors we live by. University of Chicago Press, Chicago, IL.

Langacker, R., 1991a. Concept, Image and Symbol: The Cognitive Basis of Grammar. Mouton de Gruyter.

Langacker, R. W., 1987. Foundations of cognitive grammar: Theoretical Prerequisites. Stanford University Press, Stanford, CA, vol 1.

Langacker, R. W., 1991b. Foundations of cognitive grammar: Descriptive Applications. Stanford University Press, Stanford, CA, vol 2.

Li, P., Farkaš, I., MacWhinney, B., 2004. Early lexical acquisition in a self-organizing neural network. Neural Networks 17 (8–9), 1345–1362.

Lidz, J., Gleitman, H., Gleitman, L. R., 2004. Kidz in the 'Hood: Syntactic bootstrapping and the mental lexicon. In: Hall, D. G., Waxman, S. R. (Eds.), Weaving a Lexicon. MIT Press, Cambridge, MA, pp. 603–636.

Markman, A. B., Gentner, D., 1993. Splitting the differences: A structural alignment view of similarity. Journal of Memory and Language 32, 517–535.

Markman, E., 1989. Categorization and naming in children. MIT Press, Cambridge, MA.

Markman, E., 1992. Constraints on word learning: Speculations about their origins and domain specificity. In: Gunnar, M. R., Maratsos, M. (Eds.), Modularity and constraints in language and cognition. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 59—-101.

Martin, J. H., 1991. Coding and processing of sensory information. In: Kandel, E. R., Schwartz, J. H., Jessel, T. M. (Eds.), Principles of Neural Science. Elsevier, New York, pp. 329–340.

Maturana, H. R., Varela, F. J., 1987. The Tree of Knowledge: The Biological Roots of Human Understanding. Shambhala, Boston, MA.

Medin, D. L., Altom, M. W., Edelson, S. M., Freko, D., 1982. Correlated symptoms and simulated medical classification. Journal of Experimental Psychology: Learning, Memory, and Cognition 8, 37–50.

Micko, H., 2006. Personal communication.

Minsky, M., 1975. A framework for representing knowledge. In: Winston, P. M. (Ed.), The Psychology of Computer Vision. McGraw Hill, New York, pp. 211–277.

Minsky, M., Papert, S., 1969. Perceptrons: An Introduction to Computational Geometry. MIT Press, Cambridge, Mass.

Mirolli, M., Parisi, D., 2005. Language as an aid to categorization: A neural network model of early language acquisition. In: Modelling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop. World Scientific, Singapore.

Montague, R., 1974. Formal Philosophy: Selected Papers of Richard Montague, ed. Richmond Thomason. Yale University Press, New Haven, CT.

Morris, C. W., 1938/1971. Foundations of the theory of signs. In: Writings on the Theory of Signs. Mouton, The Hague, pp. 17–74.

Narayanan, S., 1997. Knowledge-based action representations for metaphor and aspect (KARMA). Ph.D. thesis, Computer Science Division, EECS Department, University of California at Berkeley.

Návrat, P., Bieliková, M., Beňušková, L., Kapustík, I., Unger, M., 2006. Umelá inteligencia [Artificial Intelligence], 2nd Edition. STU Publishing, Bratislava.

Nehaniv, C., 2000. The making of meaning in societies: Semiotic and information-theoretic background to the evolution of communication. In: Edmonds, B., Dautenhahn, K. (Eds.), Proceedings of the AISB 2000 Symposium: Starting from Society – the Application of Social Analogies to Computational Systems. AISB, pp. 73–84.

Newell, A., 1990. Unified theories of cognition. Harvard University Press, Cambridge, MA.

Newell, A., Simon, H. A., 1976. Computer science as empirical inquiry: Symbols and search. Commun. ACM 19 (3), 113–126.

Nilsson, N. J., 1984. Shakey the robot. Tech. Rep. 323, AI Center, SRI International, Menlo Park, CA.

Nosofsky, R. M., 1984. Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition 10 (1), 104–114.

Nöth, W., 1990. Handbook of Semiotics. Indiana University Press, Bloomington, IN.

Ogden, C. K., Richards, I., 1923. The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism, 10th Edition. Routledge & Kegan Paul Ltd., London.

Oliphant, M., 1997. Formal approaches to innate and learned communication: Laying the foundation for language. Ph.D. thesis, University of California, San Diego, CA.

Oliphant, M., 1999. The learning barrier: Moving from innate to learned systems of communication. Adaptive Behavior 7 (3/4).

Orban, G. A., et al., 1995. A motion area in human visual cortex. Proc Natl. Acad. Sci. USA 92, 993–997.

Páleš, E., 1994. SAPFO, parafrázovač slovenčiny [SAPFO, Slovak paraphraser]. VEDA, Bratislava.

Parunak, H. V. D., 1996. Visualizing agent conversations: Using enhanced Dooley graphs for agent design and analysis. In: Lesser, V. (Ed.), Proceedings of the Second International Conference on Multi–Agent Systems (ICMAS'96). MIT Press, pp. 275–282.

Pecher, D., Zwaan, R. A. (Eds.), 2005. Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking. Cambridge University Press, Cambridge, U. K.

Peirce, C. S., 1931-58. Collected Writings, ed. Hartshorne, C. , Weiss, P. and Burks, A. W. Harvard University Press, Cambridge, MA.

Pfeifer, R., Scheier, C., 1999. Understanding Intelligence. MIT Press, Cambridge, MA.

Piaget, J., 1937/1955. The Child's Construction of Reality. Routledge and Kegan Paul, London, originally appeared as La construction du réel chez l'enfant. Neuchâtel, Switzerland: Delachaux et Niestlé.

Piaget, J., Inhelder, B., 1966. La Psychologie de L'enfant [The Psychology of the Child]. PUF, Paris.

Poggio, T., Girossi, F., 1990a. Networks and the best approximation property. Biol. Cybern. 63, 169–176.

Poggio, T., Girossi, F., 1990b. Networks for approximation and learning. Proc. IEEE 78 (9), 1484–1487.

Popper, M., 2007. Personal communication.

Pulvermüller, F., 1999. Words in the brain's language. Behavioral and Brain Sciences 22 (2), 253–279.

Putnam, H., 1981. Reason, Truth and History. Cambridge University Press, Cambridge.

Quine, W., 1960. Word and Object. MIT Press, Cambridge, MA.

Regier, T. P., 1992. The acquisition of lexical semantics for spatial terms: a connectionist model of perceptual categorization. Ph.D. thesis, University of California at Berkeley.

Rizolatti, G., et al., 1996. Premotor cortex and the recognition of motor actions. Cognitive Brain Research 3, 131–141.

Rogers, C., 1951. Client-centered therapy: its current practice, implications and theory. Constable, London.

Rogers, T. T., McClelland, J. L., 2004. Semantic Cognition: A Parallel Distributed Processing Approach. The MIT Press, Cambridge, MA.

Rosch, E., 1978. Principles of categorization. In: Rosch, E., Lloyd, B. (Eds.), Cognition and Categorization. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 27–48.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., Boyes-Braem, P., 1976. Basic objects in natural categories. Cognitive Psychology 8, 382–439.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386—-408.

Rosenstein, M. T., Cohen, P., Schmill, M., Atkin, M., 1997. Action representation, prediction and concepts. Working Notes of the AAAI Workshop on Robots, Softbots, Immobots: Theories of Action, Planning and Control.

Rosenstein, M. T., Cohen, P. R., 1998. Concepts from time series. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence. pp. 739–745.

Roy, D., 2005a. Grounding words in perception and action: computational insights. Trends in Cognitive Sciences 9 (8), 389–396.

Roy, D., 2005b. Semiotic schemas: a framework for grounding language in action and perception. Artificial Intelligence 167 (1–2), 170–205.

Roy, D., Hsiao, K.-Y., Mavridis, N., 2004. Mental imagery for a conversational robot. IEEE Transactions on Systems, Man, and Cybernetics, Part B 34 (3), 1374–1383.

Rubin, A., 1973. Grammar for the people: Flowcharts of SHRDLU's grammar. Tech. Rep. AIM-282, Massachusetts Institute of Technology, Cambridge, MA.

Rumelhart, D. E., McClelland, J. L., the PDP Research Group, 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. MIT Press, Cambridge.

Russell, S., Norvig, P., 1995. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ.

Rybár, J., 2005. Vývin jazyka u dieťaťa – vrodené verzus získané [Child's development of language – innate versus acquired]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 84–103.

Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), 2005. Jazyk a kognícia [Language and Cognition]. Kalligram, Bratislava.

Sapir, E., 1949. Selected Writings of Edward Sapir in Language, Culture, and Personality, ed. by D. G. Mandelbaum. University of California Press, Berkeley.

Searle, J., 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

Searle, J. R., 1980. Minds, brains, and programs. Behavioural and Brain Sciences 3, 417–457.

Sebeok, T. A., 1976. Contributions to the Doctrine of Signs. Indiana University Press, Bloomington, IN.

Šefránek, J., 2000. Inteligencia ako výpočet [Intelligence as a Computation]. IRIS, Bratislava.

Šefránek, J., 2002. Kognícia bez mentálnych procesov [Cognition without mental processes]. In: Rybár, J., Beňušková, L., Kvasnička, V. (Eds.), Kognitívne vedy. Kalligram, Bratislava, pp. 200–256.

Šefránek, J., 2005. Významy neexistujú [Meanings do not exist]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 145–187.

Šefránek, J., 2007. Personal communication.

Šefránek, J., Takáč, M., Farkaš, I., 2007. Vznik inteligencie v umelých systémoch [Origin of intelligence in artificial systems]. In: Magdolen, D. (Ed.), Hmota, život, inteligencia: Vznik. VEDA, Bratislava, in press. Extended version of this article is available online at `http://kedrigern.dcs.fmph.uniba.sk/reports` as Technical report in Informatics TR-2007-001, Comenius University, Bratislava, Slovakia.

Sejnowski, T. J., Rosenberg, C. R., 1987. Parallel networks that learn to pronounce English text. Complex Systems 1, 145–168.

Seyfarth, R., Cheney, D. L., Marler, P., 1980. Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. Science 210, 801–803.

Shastri, L., Grannes, D., Narayanan, S., Feldman, J., 1999. A connectionist encoding of schemas and reactive plans. In: Kraetzschmar, G. K., Palm, G. (Eds.), Hybrid Information Processing in Adaptive Autonomous vehicles, Lecture Notes in Computer Science. Springer-Verlag, Berlin.

Shepard, R., 1987. Toward a universal law of generalization for psychological science. Science 237, 1318–1323.

Simon, H. A., Kaplan, C. A., 1990. Foundations of cognitive science. In: Posner, M. I. (Ed.), Foundations of Cognitive Science, 2nd Edition. MIT Press, Cambridge, MA, pp. 1–47.

Siskind, J. M., 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition 61 (1–2), 1–38.

Sloan, A. P., 1978. State of the art in cognitive science. Tech. rep., Alfred P. Sloan Foundation.

Smith, A. D. M., 2001. Establishing communication systems without explicit meaning transmission. In: Kelemen, J., Sosík, P. (Eds.), Advances in Artificial Life. Proceedings of the 6th European Conference on Artificial Life ECAL 2001. Lectures Notes in Computer Science. Springer, Prague, pp. 381–390.

Smith, A. D. M., 2003a. Evolving communication through the inference of meaning. Ph.D. thesis, Theoretical and Applied Linguistics, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh.

Smith, A. D. M., 2003b. Intelligent meaning creation in a clumpy world helps communication. Artificial Life 9 (2), 559–574.

Smith, A. D. M., 2005a. The inferential transmission of language. Adaptive Behavior 13 (4), 311–324.

Smith, A. D. M., 2005b. Mutual exclusivity: Communicative success despite conceptual divergence. In: Tallerman, M. (Ed.), Language Origins: Perspectives on Evolution. Oxford University Press, pp. 372–388.

Smith, K., Smith, A. D. M., Blythe, R. A., Vogt, P., 2006. Cross-situational learning: a mathematical approach. In: Vogt, P., Sugita, Y., Tuci, E.,

Nehaniv, C. (Eds.), Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication. Springer, Berlin/Heidelberg, pp. 31–44.

Spelke, E. S., 1990. Principles of object perception. Cognitive Science 14, 29–56.

Stalnaker, R., 1981. Antiessentialism. Midwest Studies of Philosophy 4, 343–355.

Steels, L., 1997. Constructing and sharing perceptual distinctions. In: van Someren, M., Widmer, G. (Eds.), Proceedings of the European Conference on Machine Learning. Springer-Verlag, Berlin, pp. 4–13.

Steels, L., 1999. The Talking Heads Experiment. Volume 1. Words and Meanings. Laboratorium, Antwerpen.

Steels, L., 2000. Language as a complex adaptive system. In: Schoenauer, M. (Ed.), Proceedings of PPSN-VI. Springer-Verlag, Berlin, pp. 17–26.

Steels, L., Belpaeme, T., 2005. Coordinating perceptually grounded categories through language: A case study for colour. Behavioral and Brain Sciences 28 (4), 469–529.

Steels, L., Kaplan, F., 1999. Situated grounded word semantics. In: Dean, T. (Ed.), Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kauffmann, San Francisco, CA, pp. 862–867.

Steels, L., Kaplan, F., 2001a. AIBO's first words: The social learning of language and meaning. Evolution of Communication 4 (1), 3–32.

Steels, L., Kaplan, F., 2001b. Bootstrapping grounded word semantics. In: Briscoe, T. (Ed.), Linguistic Evolution Through Language Acquisition: Formal and Computational Models. Cambridge University Press, Cambridge, U. K., pp. 53–73.

Steels, L., Kaplan, F., McIntyre, A., Looveren, J. V., 2002. Crucial factors in the origins of word-meaning. In: Wray, A. (Ed.), The Transition to Language. Oxford University Press, Oxford, pp. 252–271.

Steels, L., Vogt, P., 1997. Grounding adaptive language games in robotic agents. In: Harvey, I., Husbands, P. (Eds.), Advances in Artificial Life. Proceedings of the Fourth European Conference on Artificial Life. MIT Press, Cambridge, MA, pp. 474–482.

Sussman, G., Winograd, T., 1970. Micro-planner reference manual. Tech. Rep. AIM-203, Massachusetts Institute of Technology, Cambridge, MA.

Takáč, M., 1997. Fixed point classification method for qualitative simulation. In: Costa, E., Cardoso, A. (Eds.), Progress in Artificial Intelligence: Proceedings of the Eighth Portuguese Conference on Artificial Intelligence (EPIA '97). LNAI. Springer Verlag, Berlin, pp. 255–266.

Takáč, M., 2003a. Koevolučné modelovanie vzniku jazyka [Coevolutionary modeling of language origins]. In: Kelemen, J. (Ed.), Kognice a umělý život III. Slezská univerzita, Opava, pp. 197–205.

Takáč, M., 2003b. Kultúrna dynamika v koevolúcii jazyka [Cultural dynamics in coevolution if language]. In: Sinčák, P., Kvasnička, V., Pospíchal, J., Kelemen, J., Návrat, P. (Eds.), Slovensko-České rozpravy o umelej inteligencii: Proceedings of CALCI-03. Elfa, Košice, pp. 263–268.

Takáč, M., 2003c. Kvalitatívne modelovanie a simulácia [Qualitative Modeling and Simulation]. Comenius University Press, Bratislava.

Takáč, M., 2005a. Modelovanie kultúrneho prenosu a jeho úloha v evolúcii jazyka [Models of cultural transmission and its role in language evolution]. In: Rybár, J., Kvasnička, V., Farkaš, I. (Eds.), Jazyk a kognícia. Kalligram, Bratislava, pp. 323–360.

Takáč, M., 2005b. Návrh kognitívnej architektúry pre jazykové experimenty [Cognitive architecture for language experiments]. In: Kelemen, J., Kvasnička, V., Pospíchal, J. (Eds.), Kognice a umělý život V. Slezská univerzita, Opava, pp. 549–562.

Takáč, M., 2006a. Categorization by sensory-motor interaction in artificial agents. In: Fum, D., Del Missier, F., Stocco, A. (Eds.), Proceedings of the 7th International Conference on Cognitive Modeling. Edizioni Goliardiche, Trieste, Italy, pp. 310–315.

Takáč, M., 2006b. Cognitive semantics for dynamic environments. In: Hitzler, P., Schärfe, H., Øhrstrøm, P. (Eds.), Contributions to ICCS 2006 – 14th International Conference on Conceptual Structures. Aalborg University Press, Aalborg, Denmark, pp. 202–215.

Takáč, M., 2006c. Kognitívna sémantika rozlišovacích kritérií [Cognitive semantics of identification criteria]. In: Kelemen, J., Kvasnička, V. (Eds.), Kognice a umělý život VI. Slezská univerzita, Opava, pp. 363–372.

Takáč, M., 2007a. Autonomous construction of ecologically and socially relevant semantics, Cognitive Systems Research, under review.

Takáč, M., 2007b. Kognitívna sémantika komplexných kategórií založená na rozlišovacích kritériách [Cognitive semantics of complex categories based on identification criteria]. In: Kvasnička, V., Trebatický, P., Pospíchal, J., Kelemen, J. (Eds.), Myseľ, inteligencia a život. STU Publishing, Bratislava, pp. 339–355.

Takáč, M., 2007c. Konštrukcia významov a jej dynamika v procese iterovaného učenia [Construction of meanings and its dynamics in the iterated learning process]. In: Kelemen, J., Kvasnička, V., Pospíchal, J. (Eds.), Kognice a umělý život VII. Slezská univerzita, Opava, pp. 341–347.

Takáč, M., 2007d. When meanings are not mutually exclusive: Issues in receptive field based grounded cognitive semantics. Technical Reports in Informatics TR-2007-002, Comenius University, Bratislava, Slovakia.

Talmy, L., 2000. Toward a Cognitive Semantics. MIT Press, Cambridge, MA.

Tarski, A., 1933. Pojecie prawdy w jezykach nauk dedukcyjnych [The concept of truth in the languages of the deductive sciences]. Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych 34, 13—172.

Tomasello, M., 1992. First Verbs: A Case Study of Early Grammatical Development. Cambridge University Press, Cambridge.

Tomasello, M., Farrar, J., 1986. Joint attention and early language. Child Development 57, 1454–1463.

Tschacher, W., Dauwalder, J.-P. (Eds.), 1999. The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-Organization, Embodiment, and Coordination Dynamics. Vol. 10 of Studies of Nonlinear Phenomena in Life Science. World Scientific, Singapore.

Turing, A. M., 1950. Computing machinery and intelligence. Mind 59, 433–460.

Tversky, A., 1977. Features of similarity. Psychological Review 84 (4), 327–352.

Ungerleider, L. G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D. J., Goodale, M. A., Mansfield, R. J. W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.

Vaňková, I., Nebeská, I., Římalová, L. S., Šlédrová, J., 2005. Co na srdci to na jazyku: Kapitoly z kognitivní lingvistiky [To Wear One's Heart on One's Sleeve: Chapters from Cognitive Linguistics]. Karolinum, Praha.

van Gulick, R., 1988. Consciousness, intrinsic intentionality and self-understanding machines. In: Marcel, A. J., Bisiach, E. (Eds.), Consciousness in Contemporary Science. Clarendon Press, Oxford, pp. 78–100.

Varela, F. J., Thompson, E., Rosch, E., 1991. The embodied mind. MIT Press, Cambridge, MA.

Višňovská, M., 2007. Vlastnosti a hierarchizácia kriteriálnych funkcií v konceptuálnych priestoroch [Properties and hierarchization of criterial functions in conceptual spaces]. Master's thesis, Comenius University, Bratislava.

Vogt, P., 2000. Lexicon grounding on mobile robots. Ph.D. thesis, Vrije Universiteit Brussel.

Vogt, P., 2002. The physical symbol grounding problem. Cognitive Systems Research 3 (3), 429–457.

Vogt, P., 2003a. Grounded lexicon formation without explicit reference transfer: who's talking to who? In: Banzhaf, W., Christaller, T., Ziegler, J., Dittrich, P., Kim, J. T. (Eds.), Advances in artificial life: Proceedings of the 7th European Conference on Artificial Life (ECAL03). Springer Verlag, Heidelberg, pp. 545–552.

Vogt, P., 2003b. Iterated learning and grounding: From holistic to compositional languages. In: Kirby, S. (Ed.), Proceedings of Language Evolution and Computation Workshop/Course at European Summer School in Logic, Language and Information (ESSLI). Technical University, Wien, pp. 76–86.

Vogt, P., 2005. The emergence of compositional structures in perceptually grounded language games. Artificial Intelligence 167 (1-2), 206–242.

Vogt, P., Divina, F., 2007. Social symbol grounding and language evolution. Interaction Studies 8 (1), 31–52.

von Humboldt, W., 1820/1997. Über das vergleichende Sprachstudium in Beziehung auf die verschiedenen Epochen der Sprachentwicklung [On the comparative study of language and its relation to the different periods of language development]. In: Harden, T., Farelly, D. (Eds.), Essays on Language. Peter Lang, Frankfurt am Main, pp. 65–81.

von Uexküll, J., 1909/1985. Umwelt und Innenwelt der Tiere. Julius Springer Verlag, Berlin, translated by C. J. Mellor, D. Gove. as Environment [Umwelt] and inner world of animals. In: Burghardt G. M. (ed.). Foundations of Comparative Ethology. Van Nostrand Reinhold, New York, 222–245.

von Uexküll, J., 1934/1957. A stroll through the worlds of animals and men: A picture book of invisible worlds. In: Schiller, C. H. (Ed.), Instinctive Behavior: The Development of a Modern Concept. International Universities Press, Inc., New York, pp. 5–80.

Vygotsky, L. S., 1978. Mind in Society. The Development of Higher Psychological Processes, ed. M. Cole et al. Harvard University Press, Cambridge, MA.

Warren, B., 1988. Ambiguity and vagueness in adjectives. Studia Linguistica 42 (2), 122–171.

Waxman, S. R., 2004. Everything had a name, and each name gave birth to a new thought: Links between early word-learning and conceptual organization. In: Hall, D. G., Waxman, S. R. (Eds.), Weaving a Lexicon. MIT Press, Cambridge, MA, pp. 295–335.

Waxman, S. R., Braun, I. E., 2005. Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. Cognition 95, B59–B68.

Weizenbaum, J., 1966. Eliza – a computer program for the study of natural language communication between man and machine. Communications of the ACM 9 (1), 36–45.

Weizenbaum, J., 1976. Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman & Co., New York, NY.

Wettschereck, D., Dietterich, T. G., 1992. Improving the performance of Radial Basis Function Networks by learning center locations. In: Moody, J., Hanson, S., Lippmann, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 4. Morgan Kaufmann, San Mateo, CA, pp. 1133–1140.

Whorf, B. L., 1956. Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf, ed. J. B. Carrol. MIT Press, Cambridge, MA.

Wiedermann, J., 2006. HUGO: A cognitive architecture with an incorporated world model. Tech. Rep. 966, Institute of Computer Science, Academy of Sciences of the Czech Republic.

Wiedermann, J., 2007. Nástin architektury vědomého kognitivního agenta se dvěma vnitřními modely světa [Proposal of the architecture of a conscious cognitive agent with two incorporated world models]. In: Kelemen, J., Kvasnička, V., Pospíchal, J. (Eds.), Kognice a umělý život VII. Slezská univerzita, Opava, pp. 377–383.

Winograd, T., 1971. Procedures as a representation for data in a computer program for understanding natural language. Ph.D. thesis, MIT, Cambridge, MA.

Wittgenstein, L., 1953. Philosophical Investigations. Macmillan, New York.

Ziemke, T., 1999. Rethinking grounding. In: Riegler, A., Peschl, M., von Stein, A. (Eds.), Understanding Representation in the Cognitive Sciences. Plenum Press, New York, pp. 177–190.

Ziemke, T., 2001. The construction of 'reality' in the robot: Constructivist perspectives on situated artificial intelligence and adaptive robotics. Foundations of Science, special issue on "The Impact of Radical Constructivism on Science" 6 (1–3), 163–233.