Making Sense of Words through the Eyes of a Child

A Computational Framework
for the Acquisition of Word Meanings

by
Claudius Gläser

Dissertation

submitted to the
Faculty of Technology at Bielefeld University
in partial fulfillment of the requirements for the degree of

Doktor der Ingenieurwissenschaften
(Dr.-Ing.)

October, 2011

A dissertation submitted to the Faculty of Technology at Bielefeld University for the degree of Doktor-Ingenieur (Dr.-Ing.) on October 10, 2011.

Reviewed by:

Prof. Dr.-Ing. F. Kummert	Bielefeld University, Bielefeld, Germany;
Dr.-Ing. F. Joublin	Honda Research Institute Europe GmbH, Offenbach/Main, Germany;
Prof. Dr. J. Triesch	Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe University, Frankfurt, Germany;

Accepted on May 24, 2012, on behalf of the Faculty of Technology at Bielefeld University, Germany, by the following dissertation committee:

Prof. Dr. P. Cimiano	(chairman)
Prof. Dr.-Ing. F. Kummert	(advisor)
Dr.-Ing. F. Joublin	(co-advisor)
Dr. rer. nat. T. Pfeiffer	

Claudius Gläser, »Making Sense of Words through the Eyes of a Child«

Printed on permanent paper °° ISO 9706.

Preface

The work presented in this thesis was carried out during my time as a Scientist at the Honda Research Institute Europe GmbH (HRI-EU) in Offenbach/Main, Germany. HRI-EU performs fundamental research on intelligent systems in the robotic and automotive domain since it has been inaugurated in 2003. My work was embedded into the *Child-like Acquisition of Representation and Language (CARL)* project, whose primary goal was to research methods that allow artificial systems to develop human-like language skills.

I am very grateful to many people that supported me during this exciting phase of my life. My research strongly benefited from their assistance. In the following, I would like to express my thanks to them in the language I feel most comfortable to do so.

Danksagung / Acknowledgments

An erster Stelle möchte ich mich bei Prof. Dr.-Ing. habil. Edgar Körner, Prof. Dr. rer. nat. Bernhard Sendhoff und Dipl.-Ing. Andreas Richter bedanken. Sie haben mir nicht nur die Möglichkeit gegeben, meine Forschungsinteressen am HRI-EU in die Tat umzusetzen, sondern auch den notwendigen Freiraum zur Anfertigung dieser Arbeit gewährt. Ich bin dankbar, dass ich in einem solch interessanten und gleichermaßen angenehmen Arbeitsumfeld forschen konnte. Bedanken möchte ich mich auch bei Prof. Dr.-Ing. Franz Kummert, der meine Arbeit an der Universität Bielefeld betreute, meine dortige Promotion bestmöglich unterstützte und immer ein offenes Ohr für mich hatte.

Ein besonderer Dank gilt Dr.-Ing. Frank Joublin, meinem Betreuer am HRI-EU. Er hat mich nicht nur in allen Belangen tatkräftig unterstützt, er verstand es auch, Chef und Freund zu gleich zu sein. Unser gemeinsames Büro hat sich ein ums andere Mal als Ideenschmiede entpuppt. Auch wenn ich den Tatendrang manchmal bändigen musste, waren die Diskussionen für mich immer sehr hilfreich und inspirierend. Gleichermaßen möchte ich Dr.-Ing. Martin Heckmann danken, den ich glücklicherweise als "Berater" an meiner Seite wusste. Auch wir haben viele Ideen ausgetauscht und vorangetrieben. Vor allem aber bewies er immer das richtige Gespür, geistige Höhenflüge mit kritischen Fragen jäh zu beenden, so dass ich das Ziel nie aus den Augen verlieren konnte. Von beiden, Frank und Martin, habe ich viel gelernt ... zumindest so viel, dass am Ende die Rotstifte verstummten.

Dank gilt auch Dr. Ursula Körner. Sie hat es immer geschafft, biologische Zusammenhänge so anschaulich zu vermitteln, dass sie selbst ein "Informatiker" verstehen konnte. In Bezug auf die Entwicklung von Kindern und deren Spracherwerb trifft Gleiches auf Dr.-Phil. habil. Katharina Rohlfing zu. Ich bin froh darüber, dass ich mit meiner Arbeit über den

Tellerrand schauen konnte. Sowohl Ursula als auch Katharina haben einen wesentlichen Beitrag dazu geleistet. Ich hoffe, dass diese Arbeit den Ansprüchen beider gerecht wird und bitte, etwaige Ungenauigkeiten großzügig zu überlesen.

Darüber hinaus möchte ich all meinen Kollegen am HRI-EU danken. Das gilt insbesondere für alle weiteren Mitglieder der CARL-Gruppe: Dr. rer. nat. Tobias Rodemann, Dr.-Ing. Björn Schölling, Dr.-Ing. Holger Brandl, Dr.-Ing. Xavier Domont, Dr.-Ing. Miguel Vaz, Dipl.-Ing. Irene Ayllon Clemente, Dipl.-Ing. Samuel Kevin Ngouoko und Dipl.-Ing. Rujiao Yan. Die Zusammenarbeit mit Euch hat mir sehr viel Freude bereitet.

Auch den zahlreichen Korrekturlesern dieser Arbeit gilt Dank. Dr.-Ing. Frank Joublin, Prof. Dr.-Ing. Franz Kummert, Dr.-Ing. Martin Heckmann, Dr. Ursula Körner und Dr.-Ing. Miranda Grahl haben mit ihren wertvollen Hinweisen wesentlich zum Gelingen der Arbeit beigetragen.

Abschließend möchte ich mich bei meiner Familie und meinen Freunden für deren andauernde Unterstützung bedanken. Besonders meine Eltern und Miranda haben mir immer den Rücken gestärkt und standen mir mit Rat und Tat zur Seite. So viel es mir nicht schwer, neben den Höhen auch die Tiefen einer Doktorarbeit zu meistern. Vielen Dank dafür!

Oktober, 2011 C. Gläser

Contents

Abstract

Equipping machines with the ability to understand and use natural language is a difficult task. One aspect underlying this task is the acquisition of language semantics or – to narrow the problem even more – the learning of individual word meanings. A machine, which copes with this problem, has to learn the meaning of a word based on observations of the word in different contexts. Children perform marvelously well in this task. Even though it still remains an open question how children acquire word meanings so efficiently, research in the fields of *developmental psychology* and *neurobiology* start to shed light onto some aspects of the underlying learning principles. Designing an artificial system based on such findings may consequently lead a way to overcome the restrictions of existing approaches, thereby striving towards child-like learning abilities.

In this thesis, a computational framework for the acquisition of word meanings is presented. The framework is largely inspired by findings on child development and learning. It hence establishes a link between the individual disciplines of *developmental psychology*, *neurobiology*, and *computer science*. Therefore, the thesis is structured around three central issues: Firstly, based on the abundant literature on word learning by children the different ways how children acquire word meanings are discussed. Secondly, the thesis not only discusses the respective learning processes from a phenomenological point of view, but also aims at identifying commonalities with more detailed theories on neuronal learning. More precisely, I will argue that specific neurobiological learning principles can explain the developmental patterns observed in children and, hence, may constitute the biological underpinnings of the learning processes. Lastly, based on this unified viewpoint, biologically inspired computational models for the acquisition of word meanings are presented and applied in selected word learning scenarios.

In summary, this thesis investigates a multitude of aspects that contribute to the word learning capabilities of children. It thereby promotes the view that different learning processes and biases have to be taken into account when trying to construct artificial systems with child-like learning skills. For the individual aspects, it is shown that the development of biologically inspired computational models indeed constitutes a viable approach as compared to other methods. The tight integration of the different models into a coherent overall system for word meaning acquisition, however, is necessary to finally build robots that exhibit the desired capabilities. This integration may constitute the biggest challenge future research has to overcome.

Acronyms and Abbreviations

ALISSOM	Adaptive LISSOM
ANN	Artificial Neural Network
ATL	Anterior Temporal Lobe
BDNF	Brain-derived Neurotrophic Factor
CELL	Cross-Channel Early Lexical Learning
CLS	Complementary Learning Systems
DNF	Dynamic Neural Field
EC	Entorhinal Cortex
EM	Expectation-Maximization
ERP	Event-Related Potentials
fMRI	functional Magnetic Resonance Imaging
HMM	Hidden Markov Model
IFG	Inferior Frontal Gyrus
IT	Inferior Temporal Cortex
LDA	Linear Discriminant Analysis
LEX	Lexicon of Exemplars
LISSOM	Laterally Interconnected Synergetically Self-Organizing Map
LTM	Long-Term Memory
LVQ	Learning Vector Quantization
LWPR	Locally Weighted Projection Regression
MLP	Multi-Layer Perceptron

MRAN	Minimum Resource Allocating Network
MRF	Markov Random Field
MRMI	Maximizing Renyi's Mutual Information
MTL	Medial Temporal Lobe
NGnet	Normalized Gaussian Network
NMF	Neural Modeling Field
PCA	Principal Component Analysis
pdf	Probability Density Function
PHC	Parahippocampal Cortex
pITS	posterior Inferior Temporal Sulcus
PM	Premotor Cortex
pMTG	posterior Middle Temporal Gyrus
PRC	Perirhinal Cortex
RAN	Resource Allocating Network
RAN-EKF	Resource Allocating Network using Extended Kalman Filtering
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
SNR	Signal-Noise Ratio
SOM	Self-Organizing Map
SOM	Self-Organizing Map
Spt	a region in the posterior Sylvian fissure at the parietal-temporal boundary
STG	Superior Temporal Gyrus
STM	Short-Term Memory
STS	Superior Temporal Sulcus
SVC	Support Vector Classification
SVM	Support Vector Machine

SVR	Support Vector Regression
TTX	Tetrodotoxin
TWIG	Transportable Word Intension Generator
V1	Primary Visual Cortex
WLM	Wiring Length Minimization

1 Introduction

Language is the means of getting an idea from my brain into yours without surgery.

Mark Amidon

In the recent years, robots became more and more important in our everyday life. Today, they guide us through museums, assist us while shopping, help us in the household, and in future may even be faithful companions for elderly people. These artificial agents in part exhibit capabilities that we consider to be intelligent or human-like. The more such machines enter our life, however, the more important an interaction with them will become. The most natural way of interaction between humans is communication via language. It is thus desirable that language can be used for human-robot interaction, too. This would ease the use and increase the acceptance of robots by humans. Unfortunately, the language capabilities of today's robots are far from being human-like.

Abstractly speaking, language communication can be thought of as the transmission of symbol strings as it is depicted in Fig. 1.1. To encode a message, the sender first has to choose those symbols, that refer to what he wants to communicate. The chosen symbols are subsequently assembled to a string by the use of syntax. The receiver has to decode the transmitted message. This involves the segmentation of the incoming symbol string as well as the recognition of the individual symbols, but also an retrieval of the symbols' meanings based on which the message content can be reconstructed. Importantly, the individual symbols as well as the syntax have to be shared among the sender and the receiver. Only in this way, message decoding by the receiver reveals what has been previously encoded by the sender.

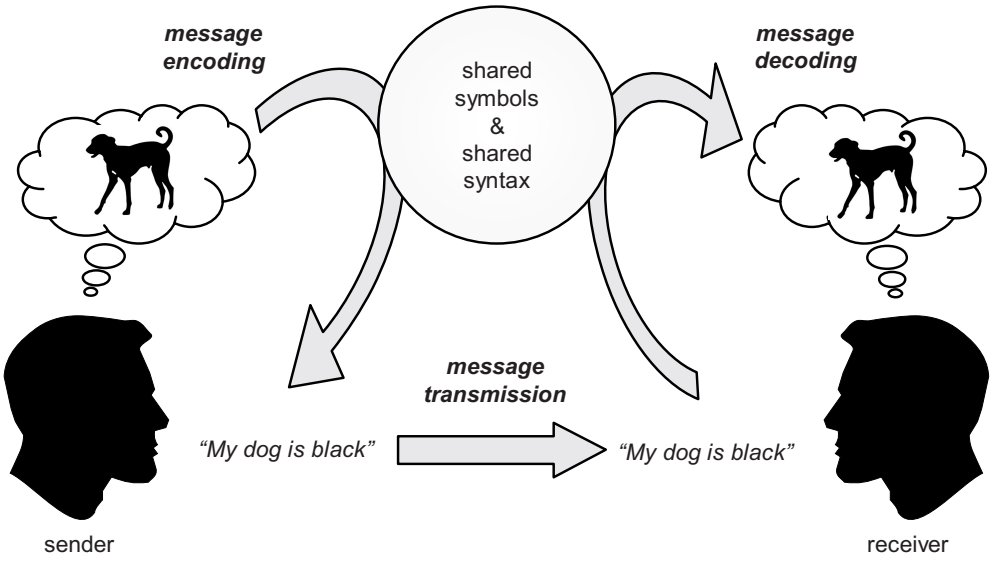


Figure 1.1.: Language communication refers to the transfer of symbolic messages. A sender encodes his thoughts in symbol strings, whereas the receiver decodes the transmitted strings to reconstruct the original message content.

In human language, symbols are words, whereas symbol strings refer to the sentences a dialog is composed of. A shared symbol lexicon therefore denotes the set of words that are known to both communication partners. The focus of the present work lies on the development of such a shared vocabulary. Other aspects involved in message encoding and decoding are out of the scope of this thesis. In other words, we assume an artificial agent to possess sufficient capabilities with respect to speech synthesis, speech segmentation, speech recognition, or syntax. This is for sure a hard restriction, since unsolved questions exist in each of these research areas. However, it is also a viable assumption as we would like to study vocabulary acquisition in isolation, i.e. without considering the problems that may arise from the other domains.

As already noted, a shared vocabulary is necessary for successful communication. To understand each other, both dialog partners need to know the communicated words and further have to associate similar meanings with them. On the contrary, a discrepancy in the vocabularies hinders conversation. The *Oxford English Dictionary* lists more than 600 000 words[1]. Even though the working vocabulary of an average English speaking person covers just a small subset of these words, it still comprises several thousand entries (Nation, 1993). For human-robot interaction this constitutes a serious problem, since it is impossible to equip robots with hand-crafted meanings for all words. This makes it difficult for humans to naturally interact with them. In fact, humans need to adapt their communication style to the language capabilities of robots.

Current robotic systems try to minimize this limitation by operating in constrained domains, e.g. as a museum guide. In such defined environments, it is often sufficient to rely on a restricted word lexicon. These systems try to spot keywords in the utterances of humans (e.g. *show* and *van Gogh*), infer what has been said by comparing the keywords

[1]Information taken from the Oxford English Dictionary's website `http://www.oed.com/` (20.09.2011)

to templates (e.g. *Show me the drawings of van Gogh, please.*), and finally react according to a defined scheme (Kopp et al., 2005). The applicability of such artificial agents is of course still limiting both with respect to the environment they may operate in as well as the way humans can interact with them.

One way that may ultimately overcome these problems is to build learning systems. In other words, systems that can extend their word repertoire during online operation and therefore are able to adapt to human interaction partners as well as changing environments. This is where word learning comes into play. Given an initially restricted or even empty lexicon, a robot may gradually increase its vocabulary size based on experience from its interactions with humans. Thereby, word learning covers two aspects: Firstly, the extraction of previously unknown symbols (words) and, secondly, the acquisition of the meanings that are associated with them. The latter aspect constitutes the focus of this thesis.

1.1. The Problem of Word Meaning Acquisition

The acquisition of word meanings is a challenging task. This is best illustrated by recapitulating a famous example stated by Quine (1960): Consider a scientist that studies a to him unknown language of a tribe. Since the scientist cannot communicate with the natives, his study is solely based on observations of what the natives are saying in which situations. At one time, he observes a native uttering the word *gavagai* while a rabbit passes by. Quine finally asked how the scientist could ever be able to infer the meaning of *gavagai*. In fact, an indefinite number of potential word meanings exist. *Gavagai* may refer to the rabbit as a whole, to any undetached part of it, to any property of the rabbit, or even to the fact that rabbits are tasty. Quine (1960) termed this problem the *indeterminacy of translation* whereas others call it *referential uncertainty* (Smith and Yu, 2008) – the uncertainty about the things a word may refer to. A word meaning obviously cannot be determined based on just a single observation of the word. Rather, a learner

Figure 1.2.: Illustration to the *gavagai example* (drawing inspired by prehistoric rock paintings).

needs further information by which wrong meaning hypotheses can be ruled out. For example, this additional information may be provided by the interaction partner, e.g. via answers to clarifying questions of the learner, or by multiple observations of the word in different contexts. In the latter case, any additional word observation would render those meaning hypotheses invalid that do not apply in the new context.

A learning robot faces a similar problem as the scientist in Quine's example. It has to determine the meaning of novel words based on its interaction with a communication partner. More precisely, symbolic descriptions (e.g. the word *gavagai*) need to be associated with representations of the environment that are internal to the robot (e.g. the concept of a rabbit that is activated by observing a rabbit). What renders the robot's word learning task even more challenging than that of the scientist is the fact that the robot not necessarily owns appropriate internal representations to which novel words can be associated. For example, if the robot initially has not been equipped with knowledge on how a rabbit looks like, the observed rabbit constitutes an unknown object for the robot. In such a case, the word *gavagai* cannot be linked to an already existing internal rabbit concept. The word rather has to be grounded in sensory input, insofar as a representation for its potential meaning, i.e. a rabbit concept, needs to be formed. Hence, word meaning acquisition additionally copes with the *symbol grounding* problem (Harnad, 1990).

In summary, for suitable human-robot interaction it is important to bridge the gap between the vocabularies of humans and robots. Building agents that are able to learn words may be the only way to achieve this, since it is impossible to predefine an appropriate word knowledge for robots. Thereby, the process of word meaning acquisition should exhibit the following characteristics:

- The vocabulary should be **unrestricted** insofar as no limits concerning lexicon size, language, task domain, or word type are placed on what can be learned.
- A **life-long** and **continuous** learning should be carried out, since any change in the environment, task domain, or interaction partner may necessitate a further extension of the vocabulary.
- Learning has to be carried out during **online operation**. At design time, it is unknown which words need to be learned by the robot (otherwise they already can be predefined). The training exemplars rather sequentially arise from the interaction with a communication partner and have to be incorporated on the fly.
- Word meanings should be acquired **fast**. This is due to the fact that the amount of training data available to an online system is limited. Humans further expect robots to show human-like capabilities which includes learning from few examples.
- Learning should **converge** towards context-independent word meanings, insofar as they finally should reflect the essential aspects of what constitutes a word referent. For example, the word *rabbit* could initially be bound to a specific instance of a white rabbit, but should finally be applicable to rabbits of any color.

In the following, it is outlined how we tried to find a solution to the aforementioned problems and requirements.

1.2. An Interdisciplinary Approach

Throughout this thesis, the development of language skills in children is taken as a role model for word learning by artificial agents. The reason for this is twofold: Firstly, the intelligent functions that today's robots offer lead to the expectation that robots behave human-like. This not only holds for the offered functions themselves, but also extends to the way robots communicate or how robots learn. Behaving like a child, rather than like an adult, obviously is an easier task and hence a more feasible research goal in a first step. Most importantly, however, children are astonishing word learners. Starting with an empty vocabulary at birth they successively increase their language skills towards adult-like performance. Thereby, children's word learning further fulfills the requirements we stated above, i.e. it is unrestricted, continuous, fast, convergent, and based on interactions with human caregivers.

This thesis hence presents an interdisciplinary approach to word meaning acquisition. More precisely, we not only describe a computational framework that can be applied in robotic agents, but also motivate the framework by findings on how children learn (see Fig. 1.3). Research in developmental psychology plays a major role in this respect. Based on experiments with children, developmental psychology is able to provide a phenomenological description of how word learning is organized in children. For example, it tells us which learning patterns are typically observed, whether there are developmental stages, or which learning processes may be involved. However, it does not give hints on how these capabilities are exactly implemented in the child's brain. For this reason, it is important to additionally consider findings from neurobiology. Establishing a link between these two disciplines finally allows us to unveil the neuronal circuits underlying children's word learning . This in turn gives important insights in how word learning could be implemented in artificial agents. The thus obtained computational models finally can

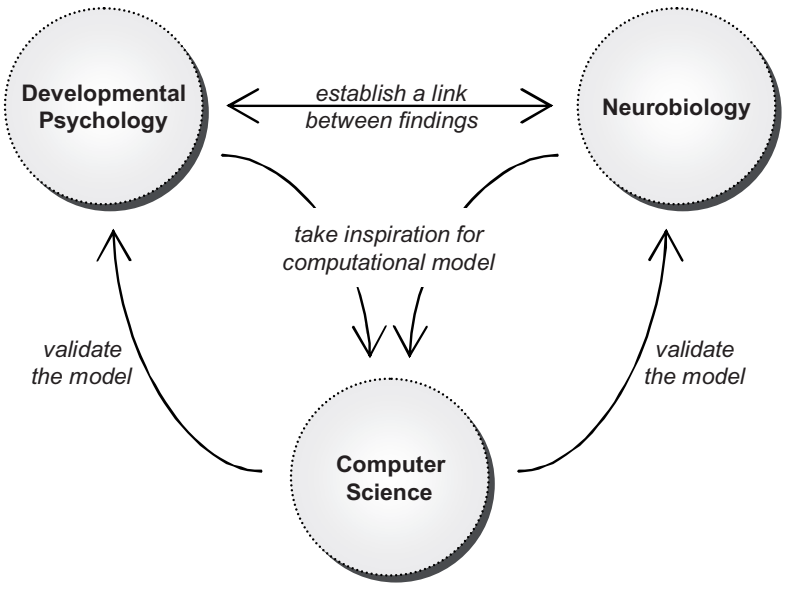


Figure 1.3.: Workflow of the interdisciplinary approach pursued in this thesis.

be validated by emulating experiments from developmental psychology and neurobiology. If the computational framework exhibits a learning behavior that is similar to that of children, the viability of the computer model would be proven.

1.3. Research Goals and Contribution of this Thesis

Even though an interdisciplinary approach to word learning seems promising, it has seldomly been pursued before. The different disciplines rather independently investigated word learning for many years. This resulted in a wealth of theories on how children learn, hypotheses on which brain areas may be involved, as well as computational models artificial agents have been equipped with. Therefore, the first goal of this thesis is to **unveil the links between findings from developmental psychology and neurobiology**. More precisely, we aim at providing a thorough literature survey based on which we first can identify the different kinds of word learning processes that may exist in children and, second, propose which neural circuitry may underlie these processes.

Once children's word learning mechanisms have been identified, they can be used to guide the development of computational models, of course. Therefore, the second goal of this thesis is to **provide biologically inspired computational models for word learning**. In detail, this comprises two aspects. Firstly, we would like to take the identified neuronal circuitry as a role model for the architecture of the computational system. This is reasonable as the circuitry suggests which kind of system components have to interact in which way to achieve an efficient word learning. Secondly, the precise implementation of the individual system components should be inspired from what is known about the computations carried out in the respective brain areas.

By building computational models that resemble their biological homologues as close as possible, we of course hope to achieve learning capabilities that are similar to that of children and hence superior to those of existing approaches. To validate this idea, the third goal of the thesis is to **thoroughly evaluate the computational models**. Firstly, this involves emulations of neurobiological experiments, insofar as it allows to check whether a particular implementation behaves alike its biological role model. Secondly, applications of the model in word learning scenarios are used to reveal whether child-like learning patterns are obtained. Finally, a comparison to existing approaches is carried out using benchmark data.

The abovementioned goals focus on the mechanisms and techniques that an artificial agent may employ to acquire word meanings efficiently. A human-robot interaction thus has not been taken into account so far. The language skills of children, however, strongly depend on the environment in which they grow up (Hart and Risley, 2003). Particularly the interaction with caregivers seems to play a key role in this respect. The fourth goal of the thesis consequently is to **estimate the influence of human-robot interaction on word learning by robots**. More precisely, we aim at investigating how different patterns of conversation between a learning agent and a tutor facilitate or hinder word meaning acquisition. This is important as it suggests how new words can be most efficiently taught to robots.

1.4. Thesis Outline

The remainder of the thesis is organized as follows. In Chapter 2 we first provide a thorough literature review. This includes an identification of language related brain areas as well as theories on how word meanings are represented in the brain. Furthermore, word learning theories stemming from the field of developmental psychology are in the scope of this chapter. They are reviewed and linked to the aforementioned neurobiological aspects, which finally allows an identification of three fundamental principles underlying children's word learning. The first two principles refer to dissociated learning systems, whereas the third principle emphasizes the role of learning constraints.

The dissociated learning systems – namely one for unsupervised and one for supervised word meaning acquisition – are in the focus of Chapter 3 and Chapter 4, respectively. For each of them, we review relevant findings from word learning experiments, suggest which neurobiological circuits may be important in this respect, and discuss the pros and cons of related computational models. We next present our biologically inspired computational models. This includes detailed algorithmic and mathematical descriptions as well as thorough evaluations of them. The latter are either done by simulating word learning scenarios or based on benchmark data. On the one hand, this allows us to emulate child experiments. On the other hand, we can assess the models' specific computational characteristics and compare their performance to those of existing approaches.

The influence of learning constraints and biases is discussed in Chapter 5. Due to the variety of constraints that have been suggested before, only one bias will be in the focus of the chapter – namely the *mutual exclusivity principle*. More precisely, we discuss the bias' computational relevance for word learning and propose a computational model for its development. An incorporation of the constraint into the previously proposed word learning framework constitutes the basis for a subsequent evaluation. We thereby investigate the influence of human-robot interaction, insofar as various conversation patterns between a tutor and a robot are evaluated with respect to their suitability for teaching new words. The chapter thus provides important insights into how a tutor should behave such that the robot can learn novel words most efficiently.

Chapter 6 finally summarizes the work and provides suggestions for future research.

2

Roots of Language Semantics

It's a strange world of language in which skating on thin ice can get you into hot water.

Franklin P. Jones (1908-1980)

Words are the fundamental building blocks of human language. They serve as symbols for aspects of the outside world as well as the inside world of a human. The use of such linguistic symbols to communicate our intents is what distinguishes humans from other species (Tomasello, 2003). But what are the origins of this seemingly human-specific language capacity? Does the human brain own a special language processing circuitry that is missing in non-human animals? Or do we (at least in part) recruit structures primary assigned to other cognitive functions? Furthermore, how much of our word learning capabilities are imprinted, i.e. relying on innate or built-in principles, and how much develop via learning from environmental input? And, finally, does one particular mechanism implement the acquisition of word meanings or do we make use of multiple strategies to ground linguistic symbols?

Cognitive scientists and biologists conducted long-lasting discussions on these questions. Some issues were answered over the last decades whereas others are still a matter of controversial debates. This chapter summarizes the current state of knowledge regarding the following aspects. It first gives an overview of the brain areas involved in language processing with a special emphasis on the *dual-stream model* and the representation of language semantics. It next provides a review of the most prominent theories on word learning and further discusses an important dissociation of symbol grounding processes, namely *concept formation* and *concept identification*. Finally, existing computational models for word meaning acquisition are reviewed and related to the aforementioned theories.

2.1. Neurobiology of Language

Knowledge about the neurobiological basis of spoken language primarily stems from behavioral studies of patients with brain lesions. Focal damage to certain brain areas results in impaired language capabilities that can be of phonological, articulatory, syntactic, or semantic nature. The first regions that have been identified this way were Broca's and Wernicke's area (Broca, 1861; Wernicke, 1874). Whereas damage to Broca's area results in impaired language articulation, damage to Wernicke's area leads to impaired language comprehension. These regions consequently have been attributed to circuits involved in speech production and speech understanding, respectively.

Later on, particularly with the advance in imaging techniques, further language-related regions have been identified (Binder et al., 1997; Price, 2000; Bookheimer, 2002). However, due to the difficulty of controlling language tasks during imaging experiments, the association of cortical areas with specific functions remains vague. Most evidence is in favor of a *dual-stream model* as proposed by Hickok and Poeppel (2004). According to the model speech is processed along two streams, a dorsal and a ventral one. As shown in Fig. 2.1, both routes initially share cortical processing in the superior temporal gyrus (STG) and the superior temporal sulcus (STS). These regions are thought to perform a spectro-temporal analysis and provide a phonological representation of speech. Afterwards, however, cortical processing diverges: A dorsal stream projects to area Spt and further to frontal regions including posterior inferior frontal gyrus (pIFG) and dorsal premotor cortex (dPM). This route hence performs a sound-to-motor mapping with area Spt providing a sensorimotor interface to a frontal articulatory network. It is noteworthy that area Spt is next to Wernicke's area, whereas the pIFG approximately resembles Broca's area. In contrast to the dorsal route, the ventral stream performs a sound-to-meaning mapping with posterior middle and posterior inferior temporal areas (pMTG & pITS) providing an interface to conceptual representations. The model further suggests that the dorsal stream is largely left-lateralized whereas processing is carried out bilaterally in the ventral stream. This is supported by the fact that more symmetric connectivity patterns seem to be beneficial for remembering semantic associations (Catani et al., 2007).

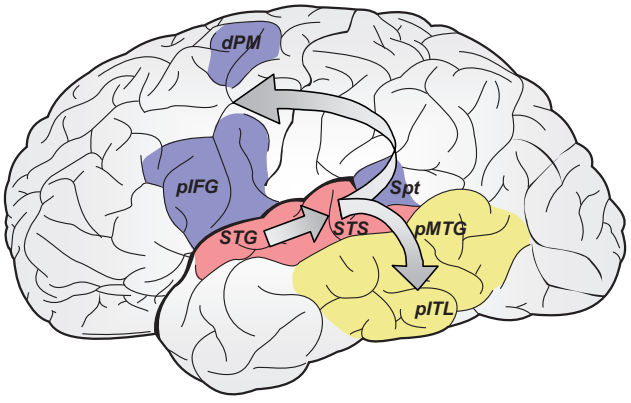


Figure 2.1.: The dual-stream model suggests that a dorsal pathway (blue) links sounds to an articulatory network whereas a ventral pathway (yellow) implements a sound-to-meaning mapping. Figure adapted from Hickok and Poeppel (2004).

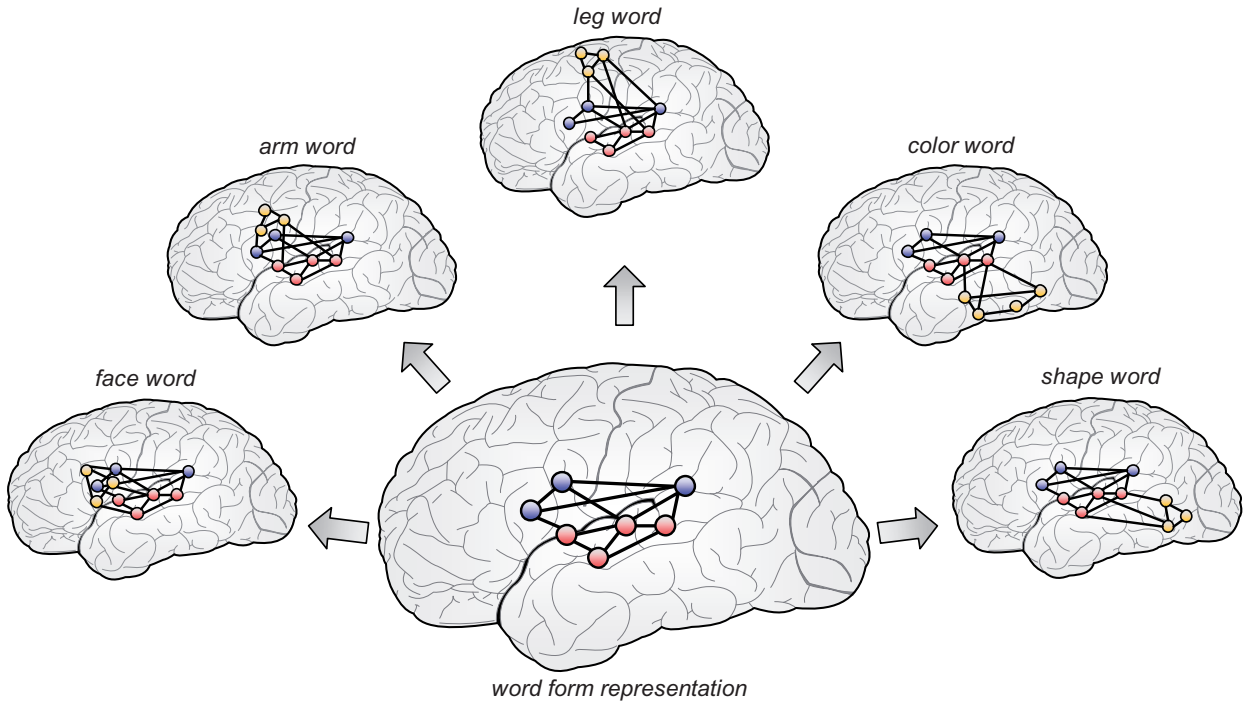


Figure 2.2.: The semantic topography model (Pulvermüller et al., 2010): Cell assemblies in the perisylvian cortex represent how words sound and how they can be articulated (center image). Word meanings are established by linking these word form representations with conceptual categories in other brain areas. Thereby, different word categories recruit different areas (satellite images).

The ventral stream is of particular importance here, since it associates word forms with their meanings. The dual-stream model though leaves open the question to what kind of conceptual representation word forms are linked. Behavioral and imaging studies suggest that the conceptual network is widely distributed. More precisely, understanding words seems to recruit different brain areas depending on the word under investigation. For example, it has been shown that the comprehension of verbs constantly activates premotor and motor regions. It could even be shown that verbs corresponding to actions of specific parts of the body (e.g. *to speak* or *to kick*) activate those motor regions that control the movements of the body parts (e.g. the mouth or the legs) (Pulvermüller et al., 2000). In the same way, color-related words evoke activity in a region of the temporal cortex that can be dissociated from the region recruited by shape-related words (Pulvermüller and Hauk, 2006). Based on these and similar findings the *neural theory of language* (Feldman and Narayanan, 2004) states that words are grounded in sensorimotor systems. The comprehension of words involves simulations of the circuits that underly the conceptual representations the words refer to. In other words, understanding a verb evokes similar activity patterns as an observation or execution of the corresponding action does (Kemmerer et al., 2008). The *semantic topography model* (see Fig. 2.2) consequently suggests that acoustic word representations in STG, STS, and pIFG are linked to category-specific conceptual representations in different brain areas (Pulvermüller et al., 2010). According to Hickok and Poeppel (2007) this coupling may be mediated by the sound-to-meaning interface of the ventral processing stream.

However, it is important to note that a concept not necessarily has to be described by only one property. For example an object concept can comprise the object's name as well as non-linguistic properties like how it looks, how it tastes, or what one can do with it. There is consequently a need to integrate information from the different conceptual categories. Compelling evidence, particularly from patients with *semantic dementia*, suggests that the anterior temporal lobe (ATL) including the more medially located parahippocampal and perirhinal cortices serve this purpose (Damasio et al., 2004; Patterson et al., 2007; Martin, 2007; Pulvermüller et al., 2010). As illustrated in Fig. 2.3, it has been proposed that the ATL acts as an amodal semantic hub by which distributed category-specific representations are bound into unique concepts. Thereby, the ATL may also be involved in grammatical constructions (Holland and Ralph, 2010) or syntactic structure building (Brennan et al., in press) – an aspect largely ignored in the present work.

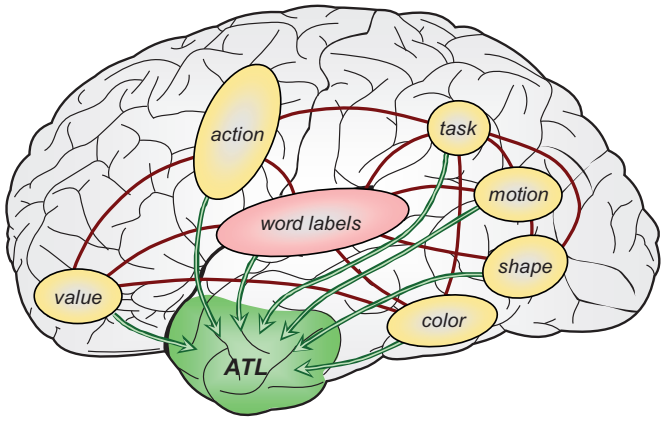


Figure 2.3.: Different words, and the conceptual categories they refer to, form a network that is distributed over large portions of the cortex. The ATL constitutes an amodal semantic hub that additionally integrates knowledge from these different areas into unique conceptual representations (e.g. object concepts). Figure adapted from Patterson et al. (2007).

Overall, the reviewed studies argue against purely language-specific brain areas in humans. It is rather proposed that language recruits sensorimotor circuits whose primary aims are other cognitive functions. From a simulation theory point of view, language enables us to communicate our intents by inducing shared cognitive states: listening to a story evokes similar activity patterns as experiencing the described situation in reality.

2.2. Word Learning Theories

The distributed network of language-related brain areas reflects the variety of concepts for which words can provide a description. It thereby also illustrates the problem that any word learner has to tackle, i.e. *referential uncertainty* – the uncertainty about the things a word may refer to. What Quine (1960) exemplarily pointed out with his *gavagei story* holds for any word learning situation. A heard word can refer to a large (or even infinite) number of aspects of the scene in which the word occurred. The learner has

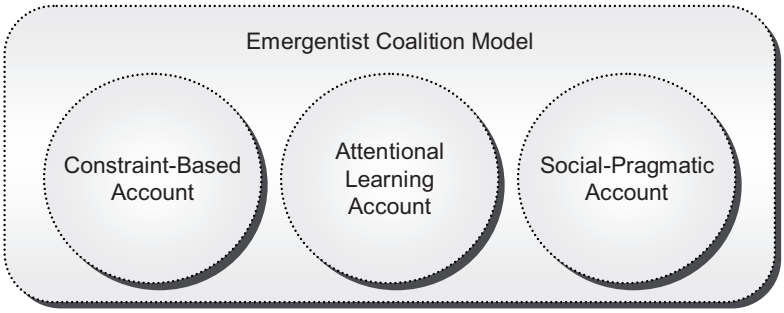


Figure 2.4.: An overview of the different word learning theories. Whereas traditional theories consider word learning in favor of one account, the Emergentist Coalition Model constitutes a hybrid theory that combines aspects of all accounts.

to pick the correct word meaning out of many possibilities. The question why children master this task so marvelously well engaged psychologists for many years. The most prominent (and often controversial) theories on word learning by children are as follows:

Constraint-Based Account (Markman, 1990): Proponents of this view emphasize the need for language-specific learning constraints that have to be innate to a child. The underlying argumentation is simple. A child cannot effectively explore the space of word meaning hypotheses, at least not at the pace typically observed in infants. Learning biases, however, can constrain the hypotheses space by pruning wrong hypotheses or facilitating attention to relevant aspects of the scene. Hence constraints can guide children's word learning. Examples for possible learning constraints are the *shape bias* (Imai et al., 1994) or the *mutual exclusivity principle* (Markman and Wachtel, 1988).

Attentional Learning Account (Smith, 1995): In strong contrast to the constraint-based account, proponents of the attentional learning account state that children do not make use of special-purpose innate structures or processes to guide their learning. They rather propose that the general principle of associative learning allows children to extract statistical regularities from the input (Plunkett, 1997). In conjunction with attentional mechanisms this general principle can develop biases which may underly children's rapid word learning (Smith et al., 1996). The theory consequently suggests that the first words are learned slowly by exploiting the statistics. However, once the first words have been learned, the acquired knowledge can be used to learn the next words more rapidly. For example, the *solid object* property can predict the relevance of the *shape* property. Selective attention on the *shape* property will finally facilitate the learning of shape-related nouns.

Social-Pragmatic Account (Bloom, 2000): Proponents of this view state that neither innate learning constraints nor general associative learning principles can fully explain the acquisition of linguistic symbols. They rather highlight the role of the social environment into which word learning is typically embedded. Thereby, particular emphasis is given to the fact that children learn words in highly structured social interactions (Tomasello,

2003). According to the theory, children do not have to tackle the problem of referential uncertainty, at least not to the extent stated by Quine (1960). Rather, caregivers restrict the hypotheses space by teaching their children in constrained and structured environments. The theory thereby assigns a pivotal role to mind reading abilities of children as aspects like *joint attention* may effectively guide the selection of the correct word referent (Bloom and Markson, 1998).

Emergentist Coalition Model (Hollich et al., 2000): As illustrated in Fig. 2.4 the aforementioned theories consider word learning in favor of one (and only one) or the other account. The emergentist coalition model breaks this barrier by combining different aspects of the previous theories into a hybrid model. Thereby, it builds on the *developmental lexical principles framework* of Golinkoff et al. (1994). At its core the model suggests that children make use of a combination of cues that guide word learning. Those cues can be social, attentional, cognitive, or linguistic in nature. It further suggests that children weight these cues differently and that cue weighting changes over the course of development. For example, perceptual saliency may be a more proficient cue than gaze following during early word learning, whereas the reverse is true at a later time. Finally, the model states that learning constraints are not innate to the child, but rather emerge and step into word learning as development progresses.

2.3. Concept Formation & Identification

As illustrated in Fig. 2.5 (a), classical theories on word learning promote a model of lexical development that differentiates between two stages. They state that concepts are first acquired via non-linguistic processes (e.g. by acting in the environment) before words become attached in order to communicate about them (Nelson, 1974). In the following the former process will be termed *concept formation* as it forms or creates new conceptual representations. In contrast, the latter process – called *concept identification* – uses already existing concepts and aims at identifying the one to which a word refers to. It consequently enriches the knowledge base by new word-concept associations but does not extent it in the form of new concepts.

When speaking about *concepts* the term *category* is often interchangeably used. This is reasonable as both terms refer to groupings of entities based on some similarity between them. As Mandler (2004) pointed out, it is however important to distinguish between *conceptual categories* and *perceptual categories*:

> "Perceptual categorization computes perceptual similarity. At least in early infancy, it does so independent of knowledge about function or kind; indeed it can occur even in the complete absence of meaningfulness. [...] Conceptual categorization computes conceptual similarity, which in the realm of objects has to do with class membership or kinds." (Mandler, 2004, p. 197)

Accordingly, key to a concept is its meaningfulness. This typically arises from its relevance to the behavior of a child, i.e. concepts provide the necessary information that allow a child to behave in a goal-directed way. Action concepts are good examples in this respect:

a child learns which actions it has to perform in order to satisfy specific needs. This suggests that *concept formation* via non-linguistic processes is primarily triggered by behavioral relevance. In fact, it could be shown that children categorize objects on the basis of their functions (Nelson et al., 2000; Booth and Waxman, 2002a; Horst et al., 2005). Once object function has been identified, children investigate object shape to unveil indicators for object function (Perone et al., 2008; Ware and Booth, 2010). Conceptual knowledge may consequently also be the precursor of attentional biases like the *shape bias* (Booth, 2006).

In addition to non-linguistic processes, it is well known that children also use labels to form conceptual categories. As illustrated in Fig. 2.5 (b), a child creates a new concept whenever it recognizes a novel word and subsequently grounds it in non-linguistic domains (e.g. color, shape, or action). Interestingly, this kind of learning supersedes a *concept identification* stage as the word is inherently associated with its conceptual category. The finding that children only accept words (but not tones) as category labels (Fulkerson et al., 2006; Ferry et al., 2010), suggests that words have a special importance due to the social relevance of language. In other words, children may think that *"if there exists a word for something, then this something has to be relevant"*. Language can consequently serve as a primary force for *concept formation*. The reverse, however, is also true (see Fig. 2.5 (c)). Once concepts have been acquired, they can guide word formation. This includes the discovery of word labels from a continous speech stream (Yeung and Werker, 2009) as well as the invention of completely new words (Gleitman and Newport, 1995). The latter may arise from the need to communicate about concepts (without having appropriate words at hand) and constitutes an important component in the evolution of new languages.

In summary, the dissociation between the possible mechanisms underlying word meaning acquisition is as follows:

1. Independent word & concept formation and subsequent concept identification
2. Concept formation driven by word labels
3. Word formation driven by concepts

The first learning mechanism can be considered to be unsupervised, insofar as word and concept formation do not have an influence on each other. In contrast, the other two mechanisms resemble supervised learning, since one modality provides supervision signals for learning in the other modality. The latter process, i.e. word formation driven by concepts, is of minor interest with regard to concept formation. Hence, it will not be considered in the rest of the thesis.

It is noteworthy that unsupervised and supervised learning mechanisms typically cannot be completely segregated. For example, it has been shown that a familiarity with objects enables children to categorize them at an earlier age (Kovack-Lesh et al., 2008). This suggests that words for objects, that have been pre-conceptualized via non-linguistic processes, are easier and faster to learn (Booth, 2009). It is known that words can alter concepts that have been previously created on the basis of non-linguistic processes (McDonough et al., 2003; Plunkett et al., 2008). A good example are spatial relations. Children acquire concepts like *above* or *below* in their first months of life via non-linguistic processes (Quinn, 2002). As it has been shown by Bowerman and Choi (2003), however,

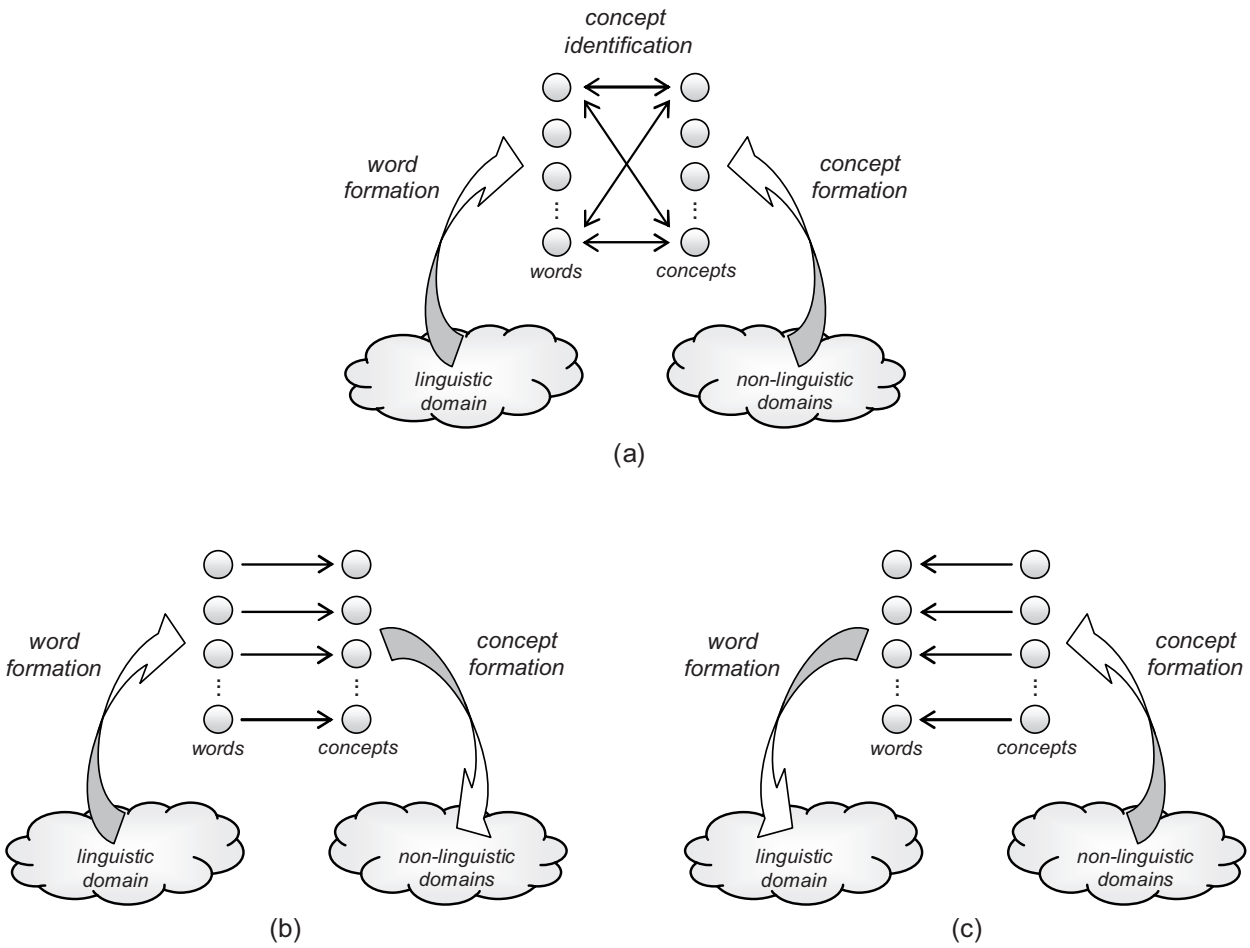


Figure 2.5.: The difference between the word learning mechanisms is illustrated. Whereas (a) depicts the process where words and concepts are independently formed and subsequently associated, (b) and (c) show that the acquisition of new words can drive the formation of corresponding concepts and vice versa.

language significantly alters these concepts afterwards, such that speakers of different languages may even have different understandings of spatiality (e.g. *containment* in English versus Korean speakers). The same is true for the color domain; people with different color vocabularies perceive colors differently (Roberson et al., 2000; Davidoff, 2001; Kay and Regier, 2006).

The findings first show that concept formation seems to attribute a higher priority to word labels than to non-linguistic processes. Secondly, the reviewed experiments provide evidence in favor of Whorfianism, i.e. language not only allows us to communicate our thoughts, but language also shapes the way we are thinking by guiding concept formation. And, finally, the strict separation between the different word learning mechanisms is questionable. The findings demonstrate that the acquisition of a particular word meaning (e.g. a color term) cannot be exclusively assigned to one or the other process. Unsupervised and supervised learning mechanisms rather seem to be heavily intertwined in many domains.

2.4. Related Computational Models

It is largely unknown how concept formation via non-linguistic processes and concept formation via word labels can be integrated into a unified model on word meaning acquisition. For this reason, existing computational approaches can be classified along these two lines. Nevertheless, the approaches differ in the level of detail they model as well as the word learning theories they adopt.

Unsupervised Concept Formation & Concept Identification

Fontanari et al. (2009) used *Neural Modeling Fields (NMFs)* to model the process of concept identification. Key to their work is the cross-situational learning approach – one of the hallmarks of associative learning theory. Thereby, cross-situational learning refers to the fact that a learner may not be able to identify the correct referent of a word from just one observation of the word. It rather suggests that hearing the word in multiple (individually ambiguous) situations enables the learner to identify its meaning. In the work of Fontanari et al. (2009) cross-situational learning appeared as follows: The learner observed scenes in which two objects were present. He additionally heard the name of one of these objects. The learner consequently had to judge to which of the two objects the word referred to. The authors showed that a batch processing of many different learning situations allows NMFs to unveil the correct word-object associations.

Xu and Tenenbaum (2007) proposed a Bayesian framework for concept identification. More precisely, the framework tries to estimate the posteriors $p(h|X)$, where X denotes a set of observed examples of a word and h a potential word meaning. Following Bayes' rule the posteriors are calculated as a product of priors $p(h)$ and likelihoods $p(X|h)$. Thereby, the priors are chosen such that different words are likely to refer to distinctive scenes. The prior thus constitutes a soft version of the mutual exclusivity principle. Furthermore, the likelihoods are evaluated in a way that favors basic-level concepts over superordinate or subordinate concepts. The model thus incorporates the taxonomic and the basic-level constraint, too. Xu and Tenenbaum consequently adopt the constraint-based theory on word learning insofar as the Bayesian framework is used to integrate evidence from the different biases. However, the model also adopts the statistical learning account, as words have to be observed in many situations to reliably estimate the required probabilities.

Yu and Ballard (2007) also proposed a probabilistic model which relates word forms to pre-established potential meanings, thereby emphasizing the role of social-pragmatic cues. More precisely, Yu and Ballard investigated the role of joint attention as a cue for referent selection as well as prosody as a cue for highlighting relevant words. To do so, the authors first transcribed a subset of the CHILDES corpus – a database containing audio and video data of mother-infant interactions. The transcription included a manual identification of objects that are simultaneously attended by the infant and the mother as well as an automatic extraction of intonation. The model was finally trained using the transcribed data, where attended objects and highlighted words were biased by assigning larger weights to them. An evaluation on two video clips showed that the incorporation of these pragmatic cues facilitated the acquisition of word-meaning associations compared to using statistical learning alone.

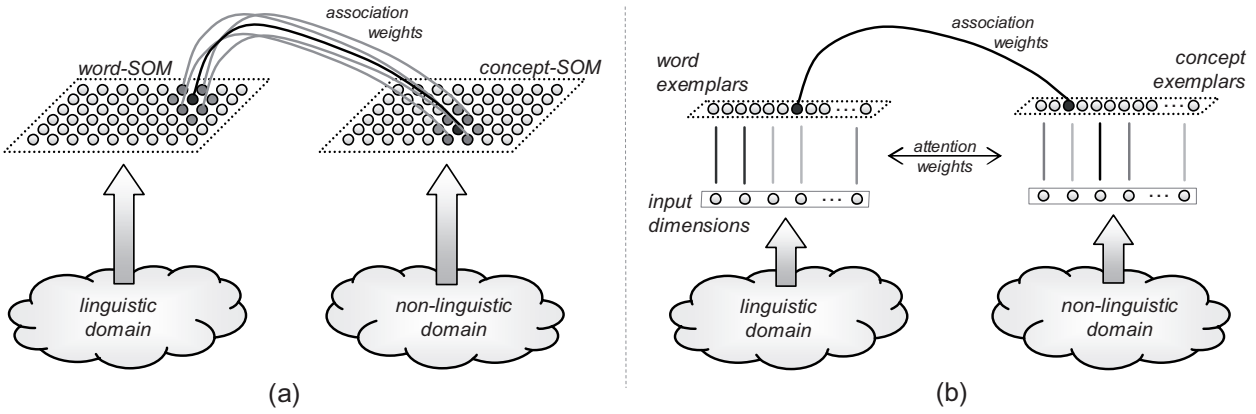


Figure 2.6.: A comparison between (a) the model of Mayor and Plunkett (2010) and (b) the LEX model of Regier (2005). Illustrations adapted from Mayor and Plunkett (2010); Regier (2005).

The aforementioned models rely on the assumption that a learner has access to a set of potential word meanings. They consequently do not tackle the problem of concept formation. In contrast, the model of Mayor and Plunkett (2010) includes both concept formation and concept identification. As illustrated in Fig. 2.6 (a), it comprises an initial batch processing of audio-visual data in which *Self-Organizing Maps (SOMs)* learn pre-lexical categories in both domains. The established categories consequently cluster entities with high acoustic or visual similarity, respectively. Concept identification is subsequently achieved via Hebbian learning of connections between both maps. In the similar vein of a bidirectional associative memory, the *Lexicon of Exemplars (LEX)* model (Regier, 2005) acquires word-meaning associations (see Fig. 2.6 (b)). However, a key difference is that LEX relies on an exemplar-based representation of pre-lexical categories instead of using SOMs to cluster them. To associate word forms and potential meanings the LEX model uses error-driven leaning of connections between the exemplars. Thereby, it further makes use of an attentional learning mechanism previously proposed by Kruschke (1992) which weights individual feature dimensions differently according to their importance (e.g. color and shape).

The model of Mayor and Plunkett (2010) as well as the model of Regier (2005) rely on a batch-processing of the data, i.e. the learner can evaluate all training data in parallel. Since word learning is obviously a continuous process in which training samples arise sequentially over time, both models only provide limited insights into how words may be acquired by children. The *CELL model* of Roy and Pentland (2002) constitutes an important advancement in this respect. As illustrated in Fig. 2.7, the CELL model uses two types of memories, a *short-term memory (STM)* and a *long-term memory (LTM)*. This memory dissociation is what supersedes the batch-processing of data, insofar as it enables a continuous processing in which observations can incrementally enter the knowledge base. More precisely, the limited-size STM acts as a first-in-first-out buffer of observations, each of them comprising a heard utterance paired with a simultaneously present object. Thereby, the object is described in terms of a shape histogram whereas the heard utterance is represented by an array of phoneme probabilities as obtained by a

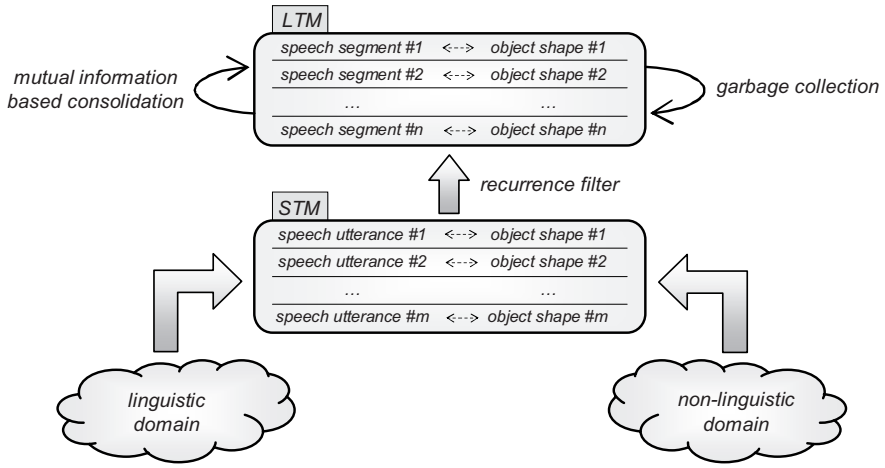


Figure 2.7.: The CELL Model. Figure adapted from Roy and Pentland (2002).

recurrent neural network. A recurrence filter subsequently processes the content of the STM by searching for sequences of phonemes which are repeatedly paired with visually similar objects. The result of this filtering are audio-visual prototypes that enter the LTM. Furthermore, the prototypes in the LTM become consolidated via a mutual information based criterion. This consolidation rejects erroneously created entries and merges similar ones which finally yields lexical items. The CELL model consequently describes one possible mechanism by which word meanings can be incrementally acquired. Thereby, it is based on recurrent coincidences of auditory and visual input. In the experiments presented by Roy and Pentland, however, multiple utterances were typically paired with the same object. Even though the visual description of the objects may slightly vary due to different views of the objects, the visual domain obviously provided the invariant information that allowed the CELL model to discover word-like units, i.e. reoccurring sequences of phonemes. The CELL model hence primarily showed how visual concepts can be used to develop word forms (cf. Fig. 2.5 (c)). Whether a similar mechanism is suitable to describe concept formation driven by word labels remains to be shown.

Supervised Concept Formation Driven by Word Labels

As already discussed, children seem to be tuned to language input, insofar as hearing words invites children to unveil their meanings (Waxman and Markow, 1995). Here, the acquisition of the word forms themselves, i.e. the segmentation of the continous speech stream into word-like units and their subsequent recognition, is not considered. If we assume a child to possess these capabilities then word meaning acquisition finally boils down to discovering commonalities among the situations in which a word occurs, since these commonalities are most likely the referent of the word. From a computational point of view this kind of learning corresponds to category formation driven by explicitly provided word labels. It is thus a supervised (or semi-supervised) process for which many computational models have been previously proposed. Nevertheless, child-like word learning possesses two important characteristics that distinguishes it from most computational approaches: Firstly, child-like learning is sequential in nature, since children

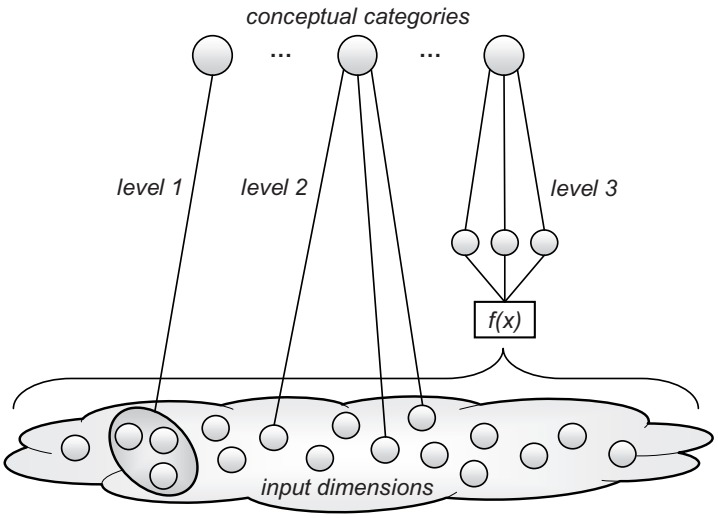


Figure 2.8.: An illustration of the three levels of (computational) referential uncertainty.

gain experience over the course of development. This rules out computational models that rely on a batch-processing of training data, i.e. those that assume that all data is known from the very beginning and can be used for learning. A valid model rather has to feature a truly incremental learning that incorporates training samples as they appear, i.e. one by one. Secondly, computational models often rely on some kind of whole-object assumption. They consider the whole scene as important and try to cluster the observations in which the same word occurred. In contrast, children's word learning is characterized by referential uncertainty. For example, children initially cannot know whether a new word refers to an object or just some property of it. They consequently have to mine those aspects of a scene that are relevant for the representation of a word's meaning.

The computational models, that will be reviewed in the following, all implement mechanisms for incremental learning. However, they differ in the level of referential uncertainty they consider. Here, it is proposed that the three levels depicted in Fig. 2.8 can be distinguished:

- Level 1: The learner already has knowledge about which aspects of a scene (in terms of sensory input dimensions) carry the relevant information for the representation of a word. This knowledge may either be innate to the system or explicitly provided during training. The computational models consequently rely on a predefined constrained input space on which category representations are built. For example, consider a system which observes objects and has to learn the meaning of color terms. A level 1 system would know, that it has to use the color of the objects instead of other dimensions (e.g. shape features) to ground the words.
- Level 2: The learner has no knowledge about the relevance of individual input dimensions. Approaches tackling this level consequently have to select the relevant feature dimensions and build categories on them. Importantly, *feature selection* thereby constrains the input space and supersedes the need for innate knowledge as compared to approaches of level 1. In the example scenario this would mean that the

system initially uses the color as well as the shape of the objects to ground the color terms. Over time, however, the system discovers that only the color dimensions are relevant for the task.

- Level 3: The relevant dimensions are hidden to the system and hence cannot be accessed directly. Models that cope with this complexity consequently have to extract the relevant information contained in the input data. This can be done by generating new feature dimensions in terms of a transformation from a set of basic input dimensions. Finally, categories can be built using the new feature space. The difference to level 2 systems consequently is that a *feature extraction* instead of a *feature selection* has to be performed. Importantly, feature extraction thereby constitutes the more general case, since it can implement any feature selection by a binary gating of input dimensions. For example, consider a system that has to learn the meaning of *dark*, but only has access to the RGB color as well as some shape features of the objects. The system cannot only select the color dimensions, since RGB color information are not sufficient to describe the *darkness* of objects. The system rather has to transform the color information into a suitable feature space (e.g. the HSV color space) and ground the word in it.

One of the representatives for level 1 approaches is the work of Steels and Kaplan (2002), who presented a system that allowed the AIBO robot to acquire object labels. Thereby, the robot has been placed in a scene in which three different objects were present. Language games (in form of questions and answers about environmental objects) served as social interaction and allowed the system to associate words with objects based on reinforcement learning (Kaplan, 1998). More precisely, the weights of an associative memory were increased or decreased based on feedback (*Yes*/*No*) on the robot's object labelling behavior. The classification as a level 1 system stems from the fact that the objects only have been represented in terms of their color – the feature that distinguished them. By doing so, knowledge about the relevant input dimension (color) has been innately given to the system which significantly reduced the difficulty of the categorization task. In a very similar way Goerick et al. (2009) taught the ASIMO robot to learn the meaning of words. In this system, however, learning was not limited to object labels, but also included words refering to object properties or actions. Observations by the robot were consequently represented using a larger number of features. However, a predefined mechanism has been used to constrain categorization, insofar as hand-crafted relevant feature spaces have been switched according to the type of word (object label, property label, or action label) that was learned.

Level 2 systems distinguish themselves from the previous approaches by including a reasoning process on the relevance of input dimensions. The *Transportable Word Intension Generator (TWIG)* of Gold et al. (2009) is one example of this class. TWIG uses decision trees to represent word meanings. Thereby, the decision nodes successively split the input space, such that the leafs of a tree represent constrained input regions which constitute the word meaning categories. More precisely, each decision node splits an input region along a hyperplane which best discriminates the referents of one word from the referents of other words. The splitting mechanism hence implements a selection of relevant feature dimensions. Even though TWIG has been demonstrated on a robotic platform, its design limits its applicability. One problem is the necessity of using predicate calculus for the

description of the environment. This puts a high burden on the system designer, since it is difficult to define sensory predicates that are suitable to describe the observations (Gold et al., 2009). A more severe problem, however, is the decision tree itself as it constructs mutually exclusive input regions. This means that an observation cannot be the referent of multiple words as it belongs to just one category (e.g. a color word, a shape word, and a label cannot refer to the same object).

Wellens et al. (2008) constructed a system in order to study the evolution of language in a population of agents (25 QRIO robots in the concrete scenario). Language games were used for social interaction between the agents. More precisely, the robots sensed their environment, described the observations they made, and finally tried to understand the descriptions given by other robots. The authors could show that this kind of interaction allowed the agents to develop a shared lexicon. In the system, the meanings of words are represented in a way akin fuzzy sets, i.e. the presence (or absence) of an object feature signals fuzzy memberships of the object with respect to the different word categories. The strengths of these feature weights represent the relevance of the different input dimensions and were learned online. By pruning features of small weights, the system hence concentrates on the most important dimensions. A drawback of the method is that it requires binary input dimensions. This means that a feature can either be present or absent. Continuous measures consequently have to be discretized via binning, which results in a rapid increase in the input dimensionality.

An additional approach for level 2 is the one of Kirstein et al. (2009). The large-scale system makes use of *Learning Vector Quantization (LVQ)* to acquire visual categories and further selects relevant input dimensions. To do so, it adopts a strategy similar to the CELL model of Roy and Pentland (2002). In detail, a *short-term memory (STM)* serves as an internal buffer for newly observed word-object pairs. A filtering mechanism subsequently processes the content of the STM and transfers the knowledge into an exemplar-based *long-term memory (LTM)*. Thereby, category-specific feature sets are created via a forward feature selection based on statistical scoring.

Surprisingly, the literature lacks level 3 approaches. To my best knowledge no computational model exists that copes with simultaneous concept formation and feature extraction. Level 3 approaches, however, are most important with respect to models of child-like learning. This is due to the fact that any life-long learning system has to cope with minimal predefined knowledge. This means that there is a need for the system to extend its internal knowledge over the course of development. This includes an extension of the conceptual network, but also an on-demand creation of new feature dimensions. A study by Schyns and Rodet (1997) provides important evidence in this respect. The authors showed that children flexibly learn new features as a consequence of categorizing objects.

3

Unsupervised Concept Formation and Word Label Mapping

Language fits over experience like a straight-jacket.

William G. Golding (1911-1993)

The acquisition of word meanings refers to a mechanism which establishes a mapping between acoustic-phonetic representations of word labels and conceptual representations that bear some kind of relevance to a child. As previously mentioned, word meaning acquisition typically cannot be attributed to one particular learning process, but rather results from a tight interaction of multiple ones. Specifically, it has been suggested in the previous chapter that the continuum of observed learning patterns may arise from a mixture of three extreme cases: (1) Words and concepts are independently formed and subsequently linked, (2) words drive the formation of concepts, or (3) concepts drive the formation of words. The latter two cases involve supervised learning, insofar as words serve as supervision signals for the formation of concepts and vice versa. This kind of learning will be addressed in the next chapter. The present chapter's focus is the independent acquisition of concepts as well as their subsequent mapping to words.

The fact that concepts can emerge without an involvement of word labels does not rule out supervised learning, of course. Other modalities such as taste or odor as well as global criteria like system performance (cf. reinforcement learning) still have an influence on how information processing is structured and therefore can 'teach' the formation of concepts. These factors, however, will not be considered here. This chapter's aim is to exploit unsupervised learning mechanisms by which conceptual representations can self-organize in a data-driven manner, i.e. solely based on the input statistic.

3.1. Self-Organization of Knowledge Representations

Unsupervised learning seems odd at first glance. Given a set of observations, a system should learn something. But in contrast to supervised learning or reinforcement learning the system is not told what to learn or how good it performs with the data. This means that the system's internal representations have to self-organize without any supervision by externally supplied criteria. Thus, the question naturally arises which objectives a system may use to guide learning. In this section, several computational goals of self-organization will be outlined and related to putative processing principles of the human brain. Finally, existing computational models are reviewed and discussed with respect to the aforementioned aspects.

3.1.1. Goals of Self-Organization

Multiple objectives of unsupervised learning can be defined (Haykin, 1998). Thereby, their individual relevance may differ depending on the concrete learning task (e.g. *feature extraction* versus *classification*). Here, the following objectives are of key interest:

- **Pattern discovery:** Unsupervised learning serves the extraction of regularities from the input data. The formation of such regular patterns (or *concepts*) is guided by their redundancy in the observations. This is what distinguishes them from pure unstructured noise (Ghahramani, 2004). Beside others, such patterns include *predictive concepts*, that give hints on possible future events, or *associative concepts*, that link multiple modalities. Here, the latter type is of particular interest as the acquisition of word meanings falls into this category.
- **Vector quantization:** The different units of a self-organized system should represent different patterns. Ideally, they should cover the whole input space such that each input pattern is appropriately represented. Moreover, this vector quantization of the input space should reflect the statistical structure of the inputs (Dayan, 1999), insofar as a fine-grained quantization is preferable for inputs that are often observed, whereas a coarse quantization is sufficient for less frequent patterns. The number of units spent to represent different inputs thus reflects their frequency of occurrence. Input space quantization and density estimation consequently is an additional objective of unsupervised learning (Duda et al., 2000).
- **Adaptivity:** Natural as well as artificial systems usually operate in instationary environments, i.e. signals that carry the information to be represented vary with time (Haykin, 1998). Thereby, signal changes may arise from the external environment (e.g. changes in lighting conditions) or from the internals of the system itself (e.g. changes in network connectivity induced by learning). An additional objective consequently is to keep track of such variations, to adapt the system to them, and thereby to stabilize the internal representations of the learned patterns.

Current computational methods for unsupervised learning already address these goals in part, but have problems in satisfying all of them simultaneously (cf. Section 3.1.3). Our approach to bridge this burden is to build bio-inspired models which take inspiration from cortical processing principles.

3.1.2. Related Principles of Cortical Processing

Self-organization of knowledge representations via unsupervised learning is one of the hallmarks of cortical development. Specifically, unsupervised learning constitutes a kind of standard paradigm employed in the brain. It is much more common than supervised learning or reinforcement learning (Dayan, 1999). Exploiting the principles of cortical processing hence is a viable approach for building computational models. The following principles may lead a way to achieve the abovementioned goals of self-organization:

- **Hebbian learning:** Connections between cortical neurons are mainly altered via Hebbian learning, i.e. based on correlated activity between individual neurons. More precisely, Hebbian learning represents the extent to which different patterns co-occur by strengthening or weakening the connections between the corresponding neurons (O'Reilly, 1998). It hence provides the possibility to extract statistical structure from the input. In the present work, Hebbian learning is of key interest as it allows to develop associative concepts, e.g. concepts that link word labels and their referents.

- **Topographic maps:** The cortex not only can be roughly divided into regions of modality-specific processing (cf. Section 2.1); the individual cortical areas further comprise maps into which the different nerve cells can be grouped (Ballard, 1997). Thereby, the term *map* refers to a layered two-dimensional plane (the cortical sheet) in which cells with similar function cluster. The visual system, for example, is organized as a hierarchy of such maps (Zeki et al., 1991). Whereas the lowest level map (V1) represents edges, the edge features are successively combined into more complex structures like curves (V2), shapes (V4), or objects (IT) at later stages of the hierarchy. A similar map-like organization can be found in other sensory modalities (Kaas et al., 1979; Schreiner, 1992; Wang et al., 1998) as well as in areas corresponding to action (Lemon, 1988) or higher cognitive function (Andersen and Buneo, 2002).

  These maps perform a vector quantization of the corresponding input spaces. Thereby, the respective input distributions are reflected, since the frequency of a particular input determines the amount of cells recruited for its representation, i.e. more frequent patterns cover larger portions of the map than less frequent patterns do (Baseler et al., 1999). Moreover, the individual cells of a map are topographically organized, insofar as nearby neurons represent similar inputs. For example, the orientation of edges smoothly varies across the V1 surface (Bonhoeffer and Grinvald, 1991). By relying on such a topographic organization, the representation of input patterns is enhanced by an additional information. This is because the (physical) positions of cell assemblies directly provide a qualitative measure on the similarity of the corresponding input patterns.

  The key principle underlying the formation of maps is *competitive learning*. This is achieved via extensive lateral interactions between cortical cells (Blakemore and Tobin, 1972). Whereas excitatory interactions are typically limited to cells within a local neighborhood, many inhibitory interactions spread over larger portions of a map (McDonald and Burkhalter, 1993). The resulting lateral inhibition between cells promotes a diversification of the cell responses. More precisely, individual cells

compete for becoming responsive to specific input patterns by hindering others to do so (by inhibiting them). Due to the fact that this is a self-enforcing process, only the 'strongest' cells will win this 'competition', whereas other cells will continue to compete for the representation of different inputs.

- **Homeostasis:** Homeostasis refers to the property of a system to regulate its internal environment to compensate for fluctuations in the external environment. It thus ensures system stability via self-maintenance of a proper operation mode. Self-regulation in the central nervous system is achieved by numerous mechanisms which act at different network scales (Marder and Goaillard, 2006). At a macroscopic level the large diversity in neuron types – particularly the heterogeneity of interneurons – plays a vital role in activity control (Santhakumar and Soltesz, 2004). Similarly, activity can be controlled at a microscopic level, e.g. by altering the strengths of synapses via homeostatic synaptic plasticity (Turrigiano and Nelson, 2004) or by varying internal neuron parameters which have an influence on the excitability of the neuron (Zhang and Linden, 2003). In the present work, we focus on these locally operating processes. A more detailed description of them will be given in Section 3.2.3 when a computational model for unsupervised map formation is introduced.

3.1.3. Existing Computational Models

Computational models for self-organizing maps already exist for a long time. Among the most popular ones are the *Kohonen maps* (Kohonen, 1982) and *Dynamic Neural Fields* (Amari, 1977). In the following, these two models as well as extensions of them will be shortly reviewed and discussed.

Kohonen Maps

Kohonen maps – as compared to Dynamic Neural Fields – model cortical map development in an abstract way. This means that Kohonen maps do not rely on detailed neurobiological mechanisms, but rather focus on the essential properties of cortical computation and implement them in a simplified fashion (Haykin, 1998). The maps are composed of units (model neurons) that are distributed on the nodes of a multi-dimensional lattice. Even though the model itself does not impose any constraint on the dimensionality of the lattice, practical applications usually rely on a 2-dimensional grid. The placement of units on the grid thereby defines the (physical) neighborhood relations between the units. Each unit i of the map possesses a set of input weights $\boldsymbol{w}_i$. This weight set serves as a *codebook vector* and defines the input pattern to which the unit is most responsive. The codebook vectors of all map units consequently vector quantize the input space. The goal of Kohonen learning therefore is to determine the units' codebook vectors such that the input space is appropriately quantized while taking the topological constraints (in terms of the spatial relations between units) into account.

To do so, learning follows a competitive-cooperative regime. For each input pattern $\boldsymbol{x}$, the map units first compete for responsibility in representing the input. Thereby, the units'

codebook vectors are compared with the input pattern. The result of this comparison is the basis for a winner-take-all decision that selects the *best-matching unit* $i(\boldsymbol{x})$. After that, the weights $\boldsymbol{w}_j$ of all units j are adapted towards the input pattern using a learning rate η and a modulation factor a_j.

$$i(\boldsymbol{x}) = \arg\min_j \parallel \boldsymbol{x} - \boldsymbol{w}_j \parallel \tag{3.1}$$

$$\boldsymbol{w}_j = \boldsymbol{w}_j + \eta \cdot a_j \cdot (\boldsymbol{x} - \boldsymbol{w}_j) \tag{3.2}$$

The factor a_j thereby controls the cooperative part of the learning. For each unit j it is chosen as a function of the (physical) distance d_{ij} between the unit and the best-matching unit i. Kohonen maps usually apply a Gaussian function $a_j = \exp(-d_{ij}^2/2\sigma^2)$ where σ determines the size of the active neighborhood. Overall, this means that the learning algorithm not only adapts the best matching unit towards the input pattern, but also those units that lay within the vicinity of the best-matching unit. In contrast, the input pattern does not affect the codebook vectors of those units that are distant from the best-matching unit.

In practice, the width σ of the active neighborhood is subject to an annealing process, i.e. starting with a large Gaussian the neighborhood successively shrinks until σ reaches a predefined minimum. As a consequence, a two-stage learning process can be observed. At the beginning, learning adapts all units towards an input pattern. This results in an initial ordering of the map during which all units roughly adapt to the input pattern distribution. Over time, however, learning becomes more specific and concentrates on units in the vicinity of the best-matching unit. This convergence phase leads to a diversification of the codebook vectors and yields the final vector quantization of the input space.

Once training converged, the quantization of the input space remains fixed. This is due to the fact that learning finally focuses only on a small local neighborhood and hence cannot significantly alter the map layout anymore. As a consequence, the adaptivity of a Kohonen map is very restricted. Some authors tried to overcome this problem by increasing or decreasing the active neighborhood on demand (Herrmann, 1995; Phaf et al., 2001). A more suitable homeostatic principle has been proposed by DeSieno (1988). In its seminal work, DeSieno introduced a conscience term to the learning algorithm. The term keeps track of how often a particular unit is selected to be the best-matching unit and finally uses this knowledge to bias the competition for subsequent input patterns. Thereby, the aim of the method is to let all units win the competition equally often. If this is the case, each unit would represent a relevant input pattern whereas the whole map would approximate the input pattern distribution. A similar idea has been pursued in the work of Sullivan and de Sa (2006). Instead of explicitly biasing the competition between map units, the authors introduced an activity-dependent scaling of the codebook vectors. By doing so, competition is implicitly biased which finally yields similar results as the method of DeSieno (1988).

Dynamic Neural Fields (DNFs)

DNF theory provides a mathematical framework by which cortical processing can be modeled at a mesoscopic level. Thereby, DNFs constitute maps in which activity is

propagated between neuron populations (Amari, 1977). Due to the variety in exhibited dynamic behavior (Coombes, 2005), neural fields have become a popular technique for modeling spatio-temporal activity flow in the brain. In detail, DNFs consider the neural tissue to be a two-dimensional plane on which neurons are distributed. The neurons are stimulated by externally applied inputs which evoke an activity within the field. Spatio-temporal response patterns are obtained by propagating activity through extensive lateral interactions between the model neurons. This dynamic spread of activity can be formally described by Amari's field equation (Amari, 1977):

$$\tau \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int w(x,x') \cdot f(u(x',t))\, dx' + S(x,t) + h. \tag{3.3}$$

Here t denotes time, $u(x,t)$ the local membrane potential of a population of neurons at position x of the cortical plane, and $S(x,t)$ the stimulus applied to this neuron population. Furthermore, neurons feature a rest potential h which is approached in absence of any other input. The monotonically increasing non-linear function relating the potential of neurons to their activities is termed f. Finally, the lateral connectivity of neurons located at position x' to neurons located at position x of the neural tissue is defined by $w(x,x')$. This interaction kernel is typically fixed and distance-dependent, i.e. $w(x,x') = w(|x-x'|)$. In most previous models a Mexican Hat connectivity is chosen. It implements an excitation between nearby neurons and an inhibition between distal ones. Hence, activity propagation within the field is competitive and can result in spatially focused regions of activity – also known as activity bubbles.

Even though DNF theory describes a general network model, a lack in understanding how neural fields can self-organize limits their applicability. Specifically, learning and adaptation have only rarely been investigated in the context of DNFs. Learning most often focuses on the synaptic weights of input projections to the neural field, thereby adapting the input-driven dynamics, but leaving the self-driven dynamics unchanged. This is due to the fact that even small learning-induced changes in the connectivity of the field can result in a significantly altered dynamic behavior of the network (Taylor, 1999; Mikhailova and Goerick, 2005). Hence, the incorporation of synaptic plasticity (e.g. via Hebbian learning) is challenging with regard to maintaining the network in stable operation modes.

The LISSOM model of Sirosh and Miikkulainen (1994) constitutes a significant advancement in this respect. In contrast to conventional models it additionally features lateral connections that undergo Hebbian plasticity. More precisely, the LISSOM model uses an interaction kernel that is initially wide and roughly Mexican Hat shaped. Subsequently, however, activity-driven learning results in a die off of synapses which sharpens and fine-tunes the kernel. The development of the interaction kernel hence can be compared to the gradually decreasing Gaussian neighborhood that is applied when training the popular Kohonen maps (Kohonen, 1982). Due to a change in the within-field connectivity, the LISSOM model is sensitive to the used parameter settings. This is why it additionally comprises an adaptation of the neuron transfer functions f that follows a predefined regime. The Adaptive LISSOM (ALISSOM) model (Law, 2009) constitutes an extension of LISSOM by means of two homeostatic mechanisms. Firstly, it uses Triesch's intrinsic plasticity model (Triesch, 2007) to adapt neuronal transfer functions and thus replaces the previously predefined regime. Secondly, ALISSOM applies activity-dependent synaptic

scaling on the afferent input connections. However, since ALISSOM does not use homeostatic principles to alter the within-field connections, it exhibits a similar parameter sensitivity as the LISSOM model. This is due to the fact that activity propagation within a neural field is largely affected by the balance between excitation and inhibition within the field. The question how a balanced lateral interaction can emerge from self-regulation is consequently an open issue.

3.2. Our Homeostatic Dynamic Neural Field Model

Even though Kohonen maps excel in simplicity and computational efficiency, we consider DNFs advantageous over them. The reasons are manifold, but mainly boil down to the dynamic nature of activity propagation within DNFs. For example, the dynamic integration of inputs enables DNFs to cope with noisy sensor data. Whereas computation in Kohonen maps relies on instantaneous sensor measurements and hence is error-prone, DNFs temporally integrate data from successive time steps. This allows DNFs to accumulate evidence which results in a dynamic competition between multiple (individually ambiguous) input pattern hypotheses. Despite the known difficulty in developing a self-organizing DNF model, we hence chose to use DNFs for unsupervised concept formation.

Our network model (Gläser and Joublin, in press), that will be presented in the following sections, differs from conventional approaches in multiple respects: (1) Similar to LISSOM the model does not make any assumption on the connectivity of the field. In other words, all synaptic weights – afferent projections to the field as well as lateral connections within the field – are plastic and change via experience-driven learning. (2) To circumvent unfavorable network behavior our model additionally employs homeostatic mechanisms. These processes operate purely locally and are based on recent findings on homeostatic principles applied in the central nervous system. This includes an intrinsic plasticity mechanism as well as homeostatic synaptic scaling. (3) In contrast to ALISSOM, however, we not only scale afferent projections to the field, but also the lateral within-field connections. By doing so, the excitation-inhibition ratio is altered. The network hence dynamically balances cooperation and competition between the model neurons in an activity-dependent manner. The following sections provide detailed information on how the individual aspects were realized in the network model.

3.2.1. Network Structure

Fig. 3.1 shows the structure of our recurrent neural network model. Similar to the Wilson-Cowan model (Wilson and Cowan, 1973) it is composed of interconnected excitatory units E and inhibitory units I. The different types of units are distributed on a layered two-dimensional grid mimicking the cortical plane. The difference to the Wilson-Cowan model lies in the connectivity between the units. Whereas Wilson and Cowan applied an all-to-all connectivity, our model consists of the following connection patterns: External input to the network – from which regular patterns should be extracted – is provided by afferent projections (w^{EXT}) to the excitatory units. Activity within the field is propagated via connections between the units. Thereby, the lateral connectivity consist

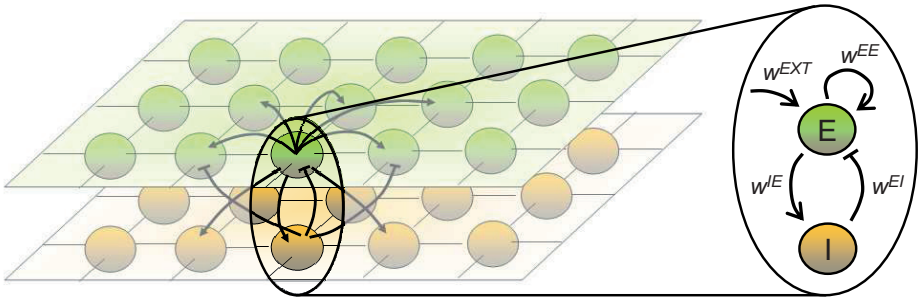


Figure 3.1.: The structure of our recurrent neural network.

of excitatory connections from E-cells to other E-cells (w^{EE}) as well as I-cells (w^{IE}). Additionally, E-cells receive inhibitory projections (w^{EI}) originating from I-cells. The direct connections between E-cells thereby implement the cooperative part of learning, whereas the coupling via I-cells serves the purpose of competitive learning.

By discretization of Amari's field equation (see Eq. (3.3)) the spatio-temporal evolution of activity within the network can be described by two differential equations. We use the variables u and v to describe the membrane potentials of the excitatory and inhibitory units, respectively. We further subset an index i to refer to the unit located at position x_i of the cortical plane:

$$\tau_E \frac{du_i}{dt} = -u_i + \sum_j g(d_{ij}) \cdot w_{ij}^{EE} \cdot f(u_j) - \sum_j w_{ij}^{EI} \cdot f(v_j) + \sum_j w_{ij}^{EXT} \cdot s_j + h^E \quad (3.4)$$

$$\tau_I \frac{dv_i}{dt} = -v_i + \sum_j g(d_{ij}) \cdot w_{ij}^{IE} \cdot f(u_j) + h^I. \quad (3.5)$$

Here, the membrane potentials are updated according to the time constants τ_E and τ_I. In absence of any input the potentials u_i and v_i approach the rest potentials h^E and h^I, respectively. Furthermore, the synaptic weight of a connection from unit j to unit i is denoted w_{ij}^* where $* \in \{EE, EI, IE, EXT\}$ describes the type of connection. The relation between the membrane potentials and the activities of units is described by the sigmoidal transfer function f which is of the form

$$f(z) = \frac{1}{1 + \exp\left(-\gamma(z - \theta)\right)}. \quad (3.6)$$

Thereby, θ is the threshold value at which a neuron exhibits an activity of 0.5, whereas γ is a gain factor that specifies the dynamic range of membrane potentials a neuron is most sensitive to.

In the update equations for the membrane potentials we additionally incorporated a modulation factor g. This factor affects the efficiency of excitatory lateral connections as a function of the distance $d_{ij} = \| x_i - x_j \|_2$ between the pre- and postsynaptic units. More precisely, we use the following function to implement an exponential decrease in

connection efficiency when distance increases:

$$g(d) = \exp\left(-\frac{d^2}{2\sigma^2}\right). \tag{3.7}$$

We consequently define that excitatory lateral connections between nearby units are more efficient than those between distant units. The term is mainly used to bootstrap the development of local representations. Its impact on network development hence is maximum at the beginning of training whereas it decreases later on. One possible interpretation of g is that of a connection probability between neurons that becomes smaller as their distance increases. Since inhibitory cells are supposed to have a broader connectivity range, they are not modulated by g in this model.

It is, however, worth noticing that the incorporation of a distance-dependent modulation factor fundamentally differs from using distance-dependent interaction kernels (as conventional approaches do). The latter implies that the synaptic weight values of lateral connections are chosen as a function of the distance between the pre- and postsynaptic units. This is not the case for our model, since we do not make any assumption on the synaptic weight values themselves. Large synaptic weight values can consequently compensate for a decrease in connection efficiency. Interestingly this also means that we withdraw the topological constraints that drive other models to develop topology preserving mappings. Hence, we hypothesize that our model produces mappings which show more topological defects than mappings developed by other approaches. For this reason, our computational model incorporates an additional and independently running process which explicitly addresses the issue of how the development of topology preserving mappings can be facilitated. We will later introduce this process in Section 3.2.4. A recent study of Hooser et al. (2005) provides evidence in favor of our model. They found orientation-sensitive cells in the primary visual cortex (V1) of a highly visual rodent, the gray squirrel. These cells are similar to those found in V1 of primates, but in contrast to primate V1 the orientation-selectivity did not smoothly vary across the cortical surface. This and other findings (Ohki and Reid, 2007) suggest that a topology preserving self-organization depends on a separate mechanism missing in rodents.

3.2.2. Hebbian Plasticity

As previously mentioned, all connections within the network model undergo experience-driven changes in synaptic strength. Thereby, the used learning regime is twofold: Firstly, it incorporates a learning rule that adapts connection weights according to the input patterns presented to the network. Secondly, it also comprises self-regulatory processes that keep the neural field in a stable state. The latter will be the focus of the following section. Here, we describe how model neurons develop appropriate representations of the input patterns via Hebbian plasticity. The learning principle stated by Hebb's rule can be shortly summarized as *cells that fire together, wire together*. In other words, if the postsynaptic cell repeatedly fires following a stimulation by the presynaptic cell, the synapse linking both cells is strengthened. To circumvent unconstrained weight growth we apply Oja's rule (Oja, 1982) which incorporates an activity-dependent leakage term:

$$\Delta w_{ij}^* \;\propto\; \eta_i \cdot \xi_j - w_{ij}^* \cdot \eta_i^2. \tag{3.8}$$

Here, w_{ij}^* is the synaptic weight, whereas η_i and ξ_j are the pre- and postsynaptic activities, respectively. Since it can be shown that Oja's rule extracts the principal component from its inputs (Oja, 1982), it constitutes a suitable learning technique for pattern discovery based on statistical regularities.

3.2.3. Homeostatic Plasticity

For neural fields, a stable operation strongly depends on balanced levels of excitation and inhibition in the network. Too much inhibition will obviously lead to vanishing activity, whereas a high level of excitation may result in runaway activity. This problem is even more severe for developing systems, since learning continuously changes network connectivity (Turrigiano and Nelson, 2000; Desai, 2003). Computational models of network development consequently have to incorporate homeostatic mechanisms to cope with these changes. In the following, we highlight recent advances in the understanding of the processes regulating neuronal activity and show how similar principles can be used within our network model.

A stable network operation constitutes itself in proper levels of network activity. Stability hence could be a consequence of activity control. In fact, studies in neuroscience provide compelling evidence for activity regulation at the level of individual neurons. For example it has been shown that neurons compensate for ongoing changes in input strength (Marder and Prinz, 2002). In these experiments, neuron cultures are placed in pharmacological substances like tetrodotoxin (TTX) which deprives the activity of the respective neurons. When this blockade is released, neurons exhibit significantly increased firing rates compared to control values. Even if the input blockade persists, cell activity gradually returns to the control level again. It could be shown that this activity regulation depends on synaptic scaling (Turrigiano et al., 1998) as well as on altering the function relating current to firing rate (Desai et al., 1999b).

Beside others (Wilhelm et al., 2009), one opinion is that synaptic scaling, i.e. the scaling of afferent projections to a neuron, is mediated by the activity-dependent release of the neurotrophin BDNF (brain-derived neurotrophic factor) (Rutherford et al., 1997). This has two important implications: Firstly, the activity of (postsynaptic) inhibitory interneurons is regulated based on the activity of the (presynaptic) excitatory cells which release the BDNF (Kokaia et al., 1993). Secondly, synaptic scaling changes the ratio between excitation and inhibition within the network. This is due to the opposite effects that BDNF has on the scaling of excitatory synapses on pyramidal neurons and interneurons, respectively (Rutherford et al., 1998). In other words, a high BDNF level weakens synapses on excitatory neurons, but strengthens those on inhibitory neurons and vice versa.

The transfer function of a neuron describes a dynamic range of input strengths to which a neuron is sensitive to. Connectivity changes induced by Hebbian learning or synaptic scaling can easily result in inputs that do not match this dynamic range. For example inputs that are too weak, such that a neuron will not fire, or inputs that are too strong, such that firing saturates. That is why an adjustment of the transfer function – so called intrinsic plasticity – is reasonable as it shifts the sensitive region such that it matches the

average input level (Desai, 2003). It is further known that this kind of self-regulation is also effected by the release of BDNF (Desai et al., 1999a).

Synaptic Scaling

In the following, we describe how we implemented a biologically inspired dynamic self-regulation in detail. Due to the activity-dependent nature of homeostasis, we first estimate the average activity level $\bar{A}_i$ of a neuron i via an integration of instantaneous activities:

$$\bar{A}_i(k) = (1 - \frac{1}{\tau_H}) \cdot \bar{A}_i(k-1) + \frac{1}{\tau_H} \cdot A_i(k) \tag{3.9}$$

Here, k is a discrete time index, $A_i(k) = f(u_i(k))$ the instantaneous activity, and τ_H defines the time scale on which integration takes place. $\bar{A}_i$ consequently can be related to intracellular calcium concentrations as they provide a correlate of a neuron's firing statistic (Berridge, 1998).

Next, we model the BDNF release of an excitatory unit i (E–cell) given its mean activity $\bar{A}_i^E$ and a target rate $\widehat{A}$ as

$$BDNF_i^E(k) = 1 + \beta_H \left(\frac{\bar{A}_i^E(k) - \widehat{A}}{\widehat{A}} \right), \tag{3.10}$$

where β_H is a homeostatic learning rate. If an E-cell's mean activity exceeds its target level, the cell's release of BDNF will be greater than 1. Conversely, the BDNF value is smaller than 1, when the cell is less active than the target level.

For the case of synaptic scaling in Kohonen-type SOMs, DeSieno (1988) previously suggested an additive scaling factor that is based on a neuron's mean firing rate. It is, however, known that multiplicative synaptic scaling is performed in the central nervous system (Turrigiano et al., 1998). This has the computationally attractive feature of leaving the relative difference in synaptic weights unchanged. A multiplicative scaling factor for SOMs, which is similar to our modeled BDNF level, has been suggested by Sullivan and de Sa (2006). However, even though our model uses the same factor for the scaling of connection weights, the way in which synaptic weights are adjusted differs fundamentally. Firstly, our learning regime combines Hebbian plasticity in form of Oja's rule with a BDNF-mediated scaling. Secondly, we further take the opposite effects of BDNF on the connections to excitatory and inhibitory cells into account. In summary, our model uses the following weight update equations:

$$w_{ij}^{EXT}(k) = \frac{w_{ij}^{EXT}(k-1) + \alpha \cdot \Delta w_{ij}^{EXT}(k)}{BDNF_i^E(k) \cdot BDNF_j^{EXT}(k)} \tag{3.11}$$

$$w_{ij}^{EE}(k) = \frac{w_{ij}^{EE}(k-1) + \alpha \cdot \Delta w_{ij}^{EE}(k)}{BDNF_i^E(k) \cdot BDNF_j^E(k)} \tag{3.12}$$

$$w_{ij}^{EI}(k) = \left[w_{ij}^{EI}(k-1) + \alpha \cdot \Delta w_{ij}^{EI}(k)\right] \cdot BDNF_i^E(k) \tag{3.13}$$

$$w_{ij}^{IE}(k) = \left[w_{ij}^{IE}(k-1) + \alpha \cdot \Delta w_{ij}^{IE}(k)\right] \cdot BDNF_j^E(k). \tag{3.14}$$

Here, α denotes a learning rate and $\Delta w_{ij}^*(k)$ the weight change according to Eq. (3.8).

Intrinsic Plasticity

In addition to synaptic scaling we model homeostatic intrinsic plasticity by altering the transfer functions of individual excitatory units. Given a sigmoidal transfer function f according to Eq. (3.6), a neuron's intrinsic excitability can be changed by dynamically adjusting the gain and threshold parameter γ and θ, respectively. In a recent work, Triesch (2007) derived an update formula for both parameters based on information theory. The difference to the mechanism applied by our model is twofold: Firstly, we restrict adaptation to the threshold parameter θ and, secondly, we express the rate of adaptation in terms of the released BDNF level:

$$\begin{aligned} \theta_i^E(k) &= \theta_i^E(k-1) + \left(BDNF_i^E(k) - 1\right) \\ &= \theta_i^E(k-1) + \beta_H \cdot \left(\frac{\bar{A}_i^E(k) - \widehat{A}}{\widehat{A}}\right). \end{aligned} \tag{3.15}$$

Homeostatic plasticity, as it is incorporated within our network model, consequently can be summarized as follows. If an excitatory neuron's average activity level exceeds its control level, the neuron releases a lot of BDNF. In turn, BDNF mediates a downscaling of synaptic weights of excitatory connections to the neuron, whereas those of inhibitory ones are upscaled. The high level of BDNF additionally triggers a decrease in the intrinsic excitability of the neuron by increasing the threshold value of its transfer function. The reverse is true when a neuron's activity level lies below its target level.

3.2.4. Topology Preservation

Even though we consider dynamic neural fields advantageous over Kohonen maps, we will discuss the issue of topology preservation also with respect to Kohonen maps. This is because our method for topology preservation is not limited to our network model; it rather can be applied to any type of SOM. When training SOMs two goals are pursued simultaneously. Firstly, SOMs perform vector quantization of the input space. They consequently strive for a minimization of the quantization error. Secondly, they incorporate topological constraints into the vector quantization process in order to develop topology preserving mappings (see Fig. 3.2 (a)). These constraints are defined in terms of fixed neighborhood relations between the map units. Unfortunately, when mapping higher-dimensional data onto the two-dimensional output space the two objectives most often cannot be simultaneously satisfied. In such cases, SOMs privilege the minimization of the quantization error at the cost of an increase in the number of topological defects within the developed mappings.

Over the past, various techniques for enhancing topology preservation during map formation have been proposed. These methods most often rely on fixed neighborhood relations between map units, but adjust the width of an active neighborhood over the course of training. This dynamic adjustment is often based on global heuristics such as a gradual decrease in the size of the active neighborhood. Alternatively, more sophisticated local measurements like input novelty (Phaf et al., 2001), topology defects (Herrmann, 1995), or the degree of local folding (Kiviluoto, 1996) can be used. Only a few approaches

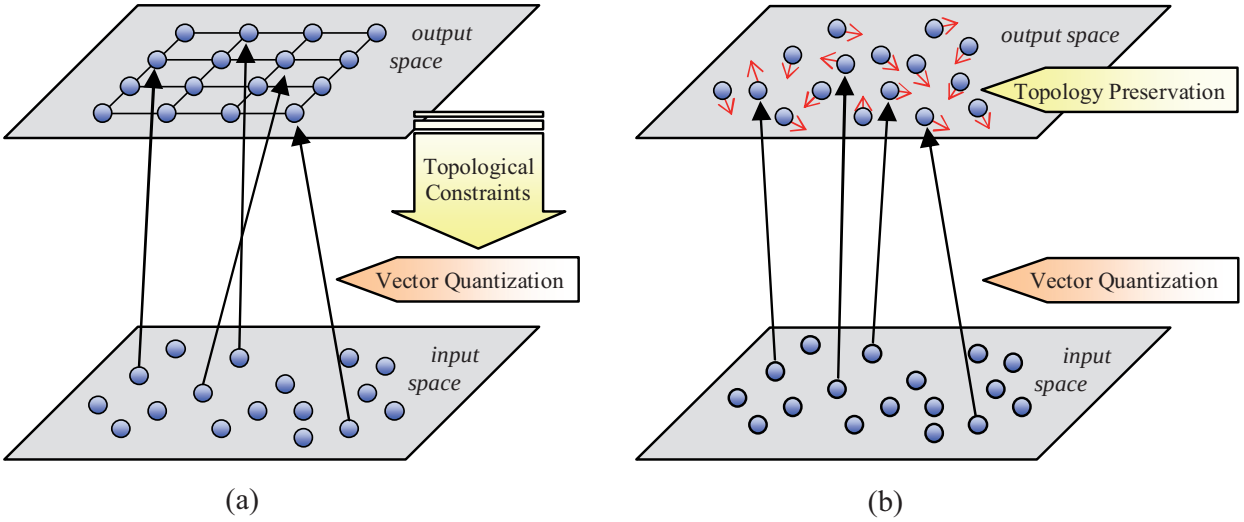


Figure 3.2.: An illustrative comparison between (a) the conventional SOM learning algorithm and (b) the proposed system for enhancing topology preservation in SOMs.

do not rely on fixed neighborhood relation. These methods rather apply a two-stage process in which vector quantization is performed first. The result of vector quantization is subsequently used to construct neighborhood relations. One example is the building of tree-like neighborhoods via a hierarchical clustering of the codebook vectors (Kirk and Zurada, 2000).

The technique we propose (Gläser et al., 2008b, 2009a) is related to this two-stage method in different respects. The first one is the release (or at least a relaxation) of the topological constraints from the process of vector quantization. As already discussed in Section 3.2.1 this is due to the fact that our network does not rely on a fixed lateral connectivity, but rather features plastic within-field connections. Similarly to the two-stage model we thus propose to incorporate an additional process which is specifically concerned with enhancing topology preservation. As it is illustrated in Fig. 3.2 (b), this means that the objective function of minimizing topological defects is no longer implicitly defined via topological constraints, but rather explicitly by a process running in parallel to the vector quantization. The key difference to the two-stage model is how this process enhances topology preservation. Here, we suggest that it changes the positions of the units in the output space. We consequently consider model neurons not to be distributed on a fixed two-dimensional grid, but rather allow them to move on the cortical plane such that they get close to other neurons with similar receptive fields.

In the following we assume neurons to be fully laterally connected, i.e. each neuron features connections to all other neurons of the SOM. Furthermore, we define the connection weight w_{ij} between two units i and j to be proportional to the similarity between the receptive fields (codebook vectors) RF_i and RF_j of the units, e.g. by

$$w_{ij} \quad \propto \quad \frac{1}{\parallel RF_i - RF_j \parallel_2}. \tag{3.16}$$

Let $d_{ij} = \parallel x_i - x_j \parallel_2$ denote the distance between two units i and j. Then we suggest to adjust the position of a unit i based on the local objective of minimizing the unit's

weighted wiring length WL_i to other units of the SOM.

$$WL_i = \sum_j w_{ij} \cdot d_{ij}^2 - \lambda \cdot \sum_j \ln(d_{ij}) \quad \longrightarrow \quad \min \tag{3.17}$$

Here, we include an additional penalization term (weighted by a factor λ) which prevents units to coincide at similar locations. Since the minimization of the distance between units with large connection weights produces a map layout where nearby units have similar receptive fields, wiring length minimization enhances topology preservation.

The movement of neurons on the cortical plane does not seem to be biologically plausible at a first glance. Even though neurogenesis (Lledo et al., 2006) – the process of continuous neuron creation from brain stem cells and their subsequent migration to target areas – describes neuron movements in the cortex, it is questionable whether neurogenesis can alter the layout of whole maps. Wiring length minimization, however, seems to be a biological principle. More precisely, it has been shown that functional brain areas as well as neuron populations within functional areas are positioned in an optimal way with respect to the achieved wiring length (Cherniak et al., 2004; Chen et al., 2006). Furthermore, a link between neuronal morphology and wiring length has been established (Chklovskii, 2004). Whether neurogenesis or structural plasticity arising from the outgrowth of axons and dendrites constitute the biological underpinning of an optimal wiring length remains an open question. Therefore, we consider our framework as a reasonable abstraction of the real biological mechanisms.

To minimize Eq. (3.17) the unit positions can be adapted using multiple optimization techniques, e.g. gradient descent or evolutionary algorithms (Fogel, 1994). Here we apply the gradient descent method insofar as the position of unit i is updated according to $\Delta x_i = -\gamma \cdot \partial\, WL_i / \partial x_i$ with

$$-\frac{\partial\, WL_i}{\partial x_i} = \sum_j 2 w_{ij} d_{ij} \cdot \frac{x_j - x_i}{d_{ij}} - \sum_j \frac{2\lambda}{d_{ij}} \cdot \frac{x_j - x_i}{d_{ij}}. \tag{3.18}$$

This formula illustrates that the map can be interpreted as an elastic network in which units exert forces on each other (see Fig. 3.3). Firstly, the lateral connections act like springs with spring constants chosen proportional to the connection weights w_{ij}. A connection between two units consequently exerts an attraction force F^+ on the units. Thereby, F^+ gets stronger when the connection weight w_{ij} or the distance d_{ij} between the units increases. Secondly, repulsion forces F^- act between the units. These forces are independent of the connection strengths. They rather solely depend on the distance d_{ij} between the units, i.e. F^- gets stronger when the distance decreases.

When wiring length minimization is applied to our network model described in Section 3.2.1, the weight values of the learned lateral connections can directly be used as connection weights w_{ij}. This is possible, because lateral connections learned via Hebbian plasticity constitute a measure for the similarity between the receptive fields of different neurons. We consequently obtain the following local objectives, where (3.19) and (3.20) hold for an

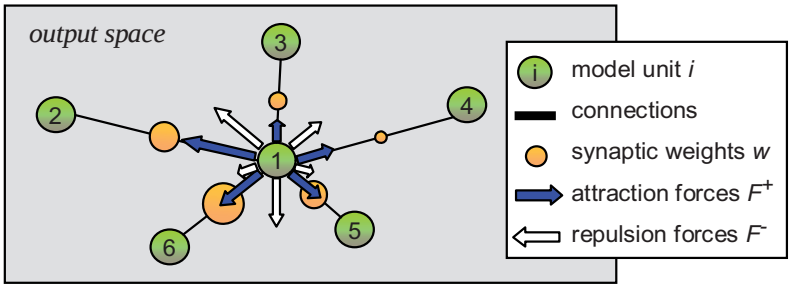


Figure 3.3.: The attraction and repulsion forces exerted on model units depend on the strengths of the connections as well as the distances between the units.

excitatory unit i and an inhibitory unit i, respectively.

$$WL_i^E = \sum_{j \in E} w_{ij}^{EE} d_{ij}^2 + \sum_{j \in I} (w_{ij}^{EI} + w_{ji}^{IE}) d_{ij}^2 - \lambda \cdot \sum_{j \in E} \ln(d_{ij}) \tag{3.19}$$

$$WL_i^I = \sum_{j \in E} (w_{ij}^{IE} + w_{ji}^{EI}) d_{ij}^2 - \lambda \cdot \sum_{j \in I} \ln(d_{ij}) \tag{3.20}$$

Here, it is important to note that our model does not incorporate repulsion forces between excitatory and inhibitory neurons. This is because both neuron types are considered to be placed on separate layers of the neural map (cf. Fig. 3.1).

3.3. Evaluation in Benchmarks

The homeostatic dynamic neural field is a general network model for unsupervised concept formation. This means that the model is not limited to linking word labels with potential referents, but rather can be applied in any domain where associative concepts have to be built. This is in accordance with what happens during child development: In many cases concepts are first built in an unsupervised fashion and subsequently linked with word labels. In Section 3.4 a particular example for such a two-stage process will be given, namely the acquisition of color categories. In this section, we first focus on the initial acquisition of concepts, i.e. without an involvement of word labels. The primary aim of this section therefore is to thoroughly evaluate the network and to unveil its computational characteristics.

We performed a series of experiments to evaluate our recurrent neural network model. In the following we first present results for a simulation where we applied the network to learn an associative mapping between artificially created multi-modal inputs. We further used the same kind of experiment to investigate how changes in stimuli strength or stimuli distribution affect neural field formation. Next, we estimated the influence of different parameter settings, i.e. different target firing rates, with respect to the developed mappings and, finally, we assessed the use of wiring length minimization for developing topology preserving mappings. The latter will be done both in the context of multi-modal association learning as well as for developing phoneme representations from continuous speech. In the following, we provide details on the obtained results.

3.3.1. Multi-Modal Association Learning

Due to their competitive processing regime, dynamic neural fields are particularly suited to learn multi-modal associations. Here we applied the model in the domain of reference frame transformation, which is a particularly important issue for robotic applications involving eye-hand coordination. Artificial agents as well as animals have to be able to flexibly transform between different frames of reference, such as body-, head-, or eye-centered coordinates (Cohen and Andersen, 2002). Agents consequently have to be equipped with an intermodal body calibration scheme which can either be innately given to the system or, more importantly, be autonomously acquired in the early stages of development (Morgan and Rochat, 1997). For the latter, the key aspect is that simultaneously present stimuli are associated in unified representations which can later be used for the transformation from one modality into another (Bahrick and Watson, 1985).

For modeling the body calibration process we restricted ourselves to a one-dimensional eye-hand coordination task. Thereby, a simulated agent performs random hand and eye movements, i.e. target gaze and hand positions are randomly chosen whereas a linear dynamic model produces smooth transitions between successive target positions. This kind of behavior emulates the self-exploratory actions that can be observed in early infancy. The agent further perceives its resulting gaze and hand positions in different reference frames. In the experiment we use three stimuli s^1,s^2,s^3 with $s^1, s^2 \in [-1, 1]$ and $s^3 = s^1 - s^2$, where s^1 and s^2 mimic the gaze and hand position in a body-centered reference frame, respectively, as well as s^3 representing the hand position in eye-centered coordinates. A specific body state consequently yields stimuli that produce individually ambiguous activities in each input modality. Their combination, however, provides a unique description of the body state. In our setup each of the stimuli is represented by a population of 21 neurons with partly overlapping Gaussian-shaped receptive fields (see Fig. 3.4). For s^1 and s^2 the receptive fields have a standard deviation of 0.1 and their centers were uniformly placed in the interval $[-1, 1]$. The centers of the receptive fields for s^3 have been uniformly sampled from the inverse of the cumulative density function of the normal distribution with standard deviation 0.4. Thereby, the receptive fields have a standard deviation that is half the distance to their nearest neighbor.

To learn associations between the different reference frames we use the following system setup: The network is composed of 100 excitatory units and 100 inhibitory units, both arranged on a 10x10 grid. The lateral connections within the field are initialized with uniform weight values, whereas the weights of afferent projections to the field are initialized with small random values. The time constants of the model are set to $\tau_E = \tau_I = 5$ and $\tau_H = 10^4$. The large homeostatic time constant τ_H ensures that average firing rates are based on a long time interval and not affected by moment-to-moment fluctuations in activity. We further use learning rates of $\alpha = 10^{-3}$ as well as $\beta_H = 10^{-4}$, i.e. Hebbian plasticity is faster than homeostatic plasticity. Finally, we apply a target activity level of $\widehat{A} = 0.05$. In the present experiment, we do not perform wiring length minimization and learning is carried out at each time step, i.e. we do not make any assumption on when learning takes place. This is in contrast to the LISSOM models (Sirosh and Miikkulainen, 1994; Law, 2009). There the synaptic weights were changed only after the network settled into a stable state following stimuli presentation.

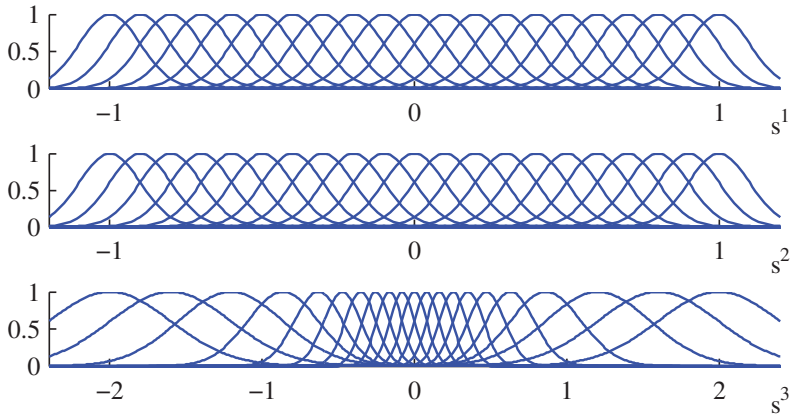


Figure 3.4.: The receptive fields of the neurons used for coding the gaze position in body-centered coordinates (s^1), the hand position in body-centered coordinates (s^2), and the hand-position in an eye-centered reference frame (s^3).

When applying the network to the sequentially arising stimuli, different phases can be observed over the course of development. Initially the model units cooperate via lateral excitation such that the whole field grossly adapts to the input pattern distribution. However, afterwards an increased lateral inhibition implements a competition between the model units. More precisely, the excitatory units compete for the representation of different input patterns. This competitive learning facilitates a diversification of the units and results in a specialization of the units to distinct input patterns. After several input patterns have been presented, we fixed the network weights and calculated the receptive fields of the excitatory units. Therefore, we applied different stimuli combinations and recorded the units' activities after the field activity settled into a stable state. The resulting receptive fields are depicted in Fig. 3.5 (a) where we plot the response pattern of each excitatory unit to different combinations of s^1 and s^2. As can be seen, each neuron specializes to a particular combination of the stimuli. We further calculated the center of masses of the receptive fields. By doing so, we obtain the positions of the neurons in the input space (the codebook vectors). Fig. 3.5 (b) illustrates that the neurons cover the whole input space (the s^1-s^2-s^3-plane), i.e. each input pattern is adequately represented by the neural field.

Lastly, we investigated whether the incorporated homeostatic mechanisms drive the individual neurons towards some target firing rate. Therefore, we recorded how the average activity levels of all excitatory neurons develop over time. This has been done for two simulations using target firing rates of $\widehat{A} = 0.05$ and $\widehat{A} = 0.1$, respectively. Fig. 3.6 plots the medians of the resulting activity levels. The regions around the medians depict the upper and lower quartiles of the activity level distributions. The plot illustrates that the neurons' average activities quickly raise towards the specified target firing rates. Additionally, we could previously show that the overall activity within the field approaches a level which is proportional to the target firing rate (Gläser et al., 2008c). In summary this shows that the applied locally operating mechanisms are suited not only to regulate an individual neurons activity, but also to regulate the activity within a population of neurons.

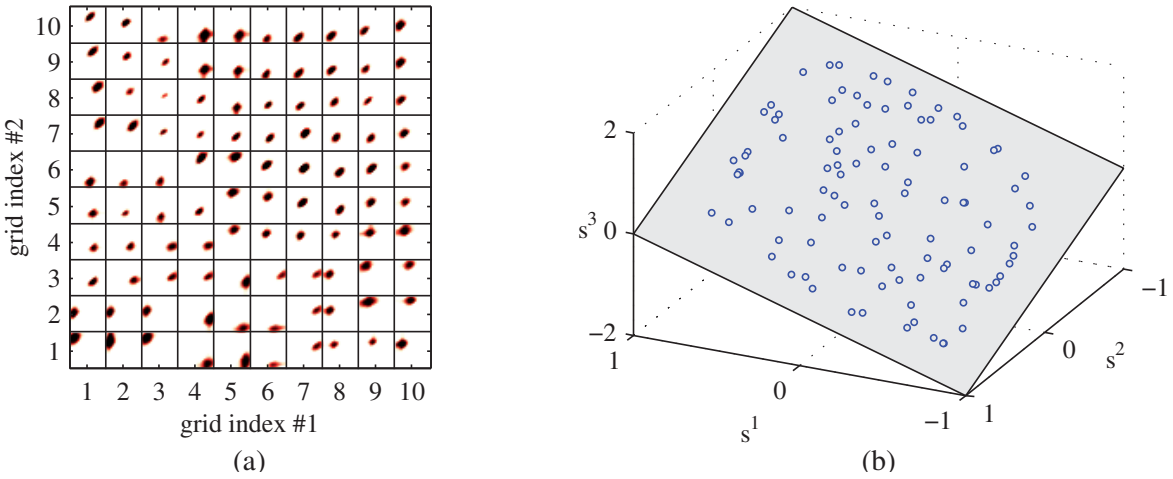


Figure 3.5.: (a) The receptive fields of all excitatory units are shown. Here, each subimage corresponds to the response pattern of a particular neuron to different combinations of s^1 (x-coordinate) and s^2 (y-coordinate). Dark colors represent strong neuron activities whereas light colors correspond to weak responses. (b) The positions of the excitatory neurons in the input space as obtained by calculating the center of masses of their receptive fields.

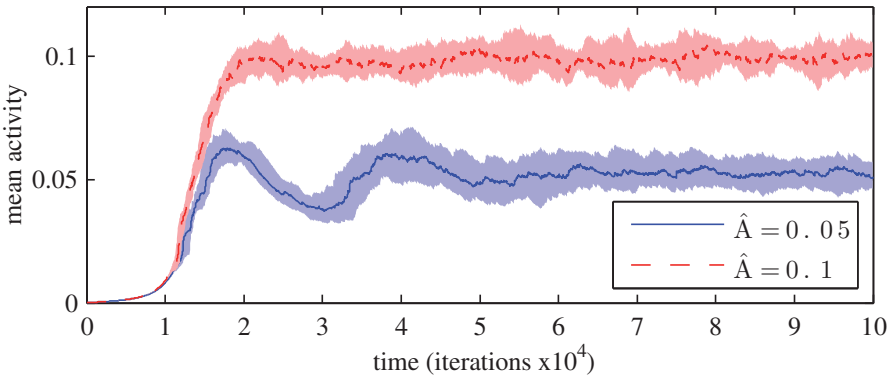


Figure 3.6.: The median of the average activity levels of all excitatory neurons is plotted for two simulations using a target firing rate of $\widehat{A} = 0.05$ and $\widehat{A} = 0.1$, respectively. Regions around the medians depict the upper and lower quartiles of the respective activity level distributions.

Effect of Changes in Stimuli Strength

Once neurons are equipped with the ability to regulate their activity, the question arises whether the same homeostatic processes are suited to adapt the dynamic neural field to long-lasting changes in the input stimuli. This includes changes in stimuli strength as well as changes in the input pattern distribution. Here, we first concentrate on the former aspect. The latter will be investigated afterwards.

To test the ability of our network to adapt to changing input strengths we simulated a biological experiment in which the input to a neuron is blocked (Marder and Prinz, 2002) (see Section 3.2.3 for a description of the experiment). More precisely, we modeled the

blockade of excitatory inputs by an attenuation of the input amplitude to 20% compared to normal operation. After a while this blockade is released again. We recorded the evolution of the BDNF level $BDNF_1^E$ as well as the transfer function threshold θ_1^E of the respective neuron. The corresponding plots are shown in Fig. 3.7. Here, time $t = 0$ denotes the onset of stimulus depression, whereas at time $t = 100$ the blockade is released. At both times, we further recorded the responses of the neuron to a specific input pattern, once using normal inputs strength and once using an input strength depressed to 20% of normal operation. The inset plots at time $t = 0$ show that the presentation of the input pattern without blockade produces a stable and large response of the neuron. In contrast, the depressed input pattern is too weak to produce a significant increase in the neuron's potential such that the neuron remains inactive. Due to this behavior, the neuron's mean activity level will decrease after the onset of blockade at $t = 0$ (not shown). As a result, the neuron compensates for this change in a similar way as biological neurons do: It decreases its BDNF level, which results in an upscaling of excitatory, but a downscaling of inhibitory synapses. It further decreases its threshold and thus changes its transfer function towards higher excitability. The result of this regulation is depicted by the inset plots at time $t = 100$. The depressed input pattern now induces the same response of the neuron as the normal pattern did at time $t = 0$. However, if we release the blockade, i.e. present the undepressed input pattern, then the neuron shows a much larger sensitivity to the pattern. This reflects itself in the significantly increased potential evoked by the input and consequently a faster and prolonged response of the neuron.

In summary these results show that the homeostatic mechanisms enable individual neurons to compensate for changes in the strengths of their inputs. From a computational point of view this is advantageous as it allows the network to cope with inputs which may continuously change their amplitude over a long time scale.

Effect of Changes in the Input Distribution

Next, we investigated whether our network model is able to adapt an already developed mapping to a changed input pattern distribution. Biological neurons can cope with such changes (Joublin et al., 1996). They are even able to adapt to sudden and significant changes such as those following the amputation of a limb (Halligan et al., 1993). For example it has been shown that digit amputation in raccoon forces affected neurons in primary somato-sensory cortex to reorganize their receptive fields (Foeller and Feldman, 2004). More precisely, neurons that become silent after amputation (due to missing input) subsequently expand their receptive field to large regions of adjacent digits or the palm and finally shrink them again such that the neuron becomes selectively responsive to inputs stemming from the new receptive field. To test our computational model we perform an experiment of a similar kind. Therefore, we first let the network develop a mapping, but subsequently we significantly change the input pattern distribution at regular time intervals.

The results of this simulation are depicted in Fig. 3.8. Here, the network initially has been trained with uniformly sampled input stimuli, i.e. $s^1 \sim \mathcal{U}(-1, +1)$, $s^2 \sim \mathcal{U}(-1, +1)$,

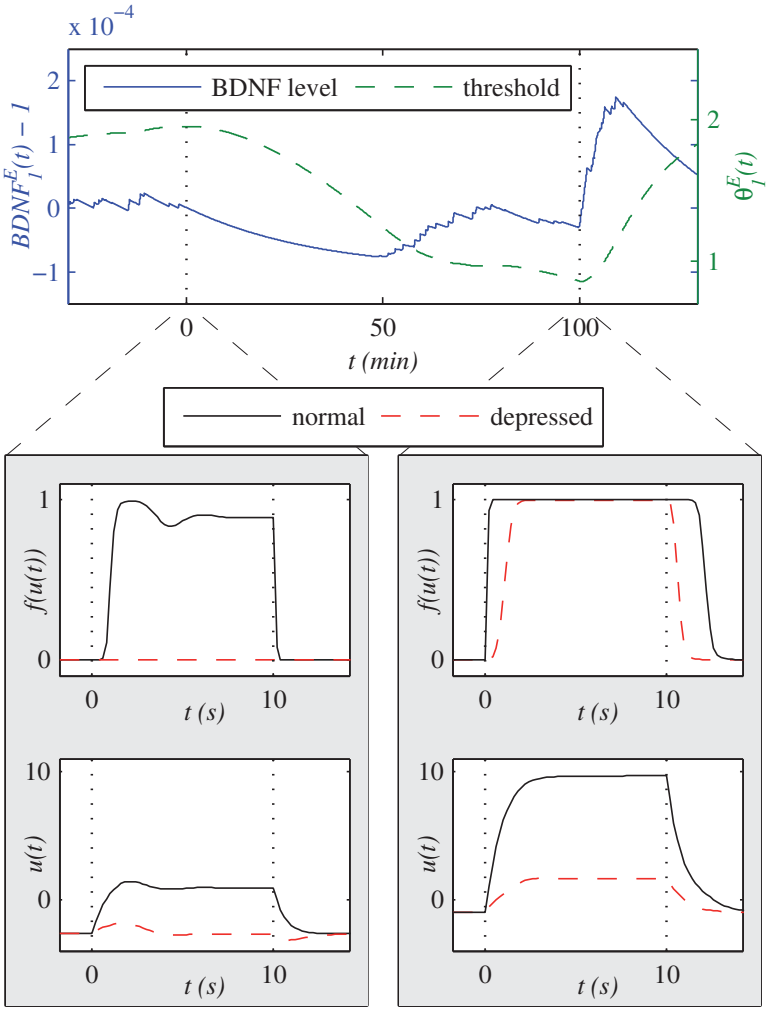


Figure 3.7.: The evolution of a neuron's BDNF level and transfer function threshold following an input blockade at $t = 0$ as well as a release of the blockade at $t = 100$. The blockade has been modeled by an attenuation of input strengths to 20% of normal operation. The insets shows the response of the neuron to the normal as well as the depressed input pattern when they are presented at times $t = 0$ and $t = 100$, respectively.

and $s^3 = s^1 - s^2$. As shown in the inset at time $t = 0$, the network developed a mapping where the receptive fields of the excitatory neurons are nicely distributed in the input space. At time $t = 0$ we applied the first change in the input pattern distribution. At subsequent time steps, stimuli with $s^3 \sim \mathcal{N}(0, 0.09)$ have been sampled and presented to the network. Neurons, which previously developed a receptive field corresponding to large absolute values of s^3, consequently do not receive inputs anymore. In contrast, those neurons, with receptive fields already lying close to $s^3 = 0$, now become activated by more input patterns than before the change. This is reflected in the neurons' average activity level distribution which is plotted at the bottom panel of Fig. 3.8. Since most neurons do not become activated anymore, the median as well as the lower quartile of the distribution decrease during the time steps following $t = 0$. However, the change in the input pattern distribution also increases the average activity level of some neurons, such that the upper quartile of the distribution rises. The homeostatic processes consequently

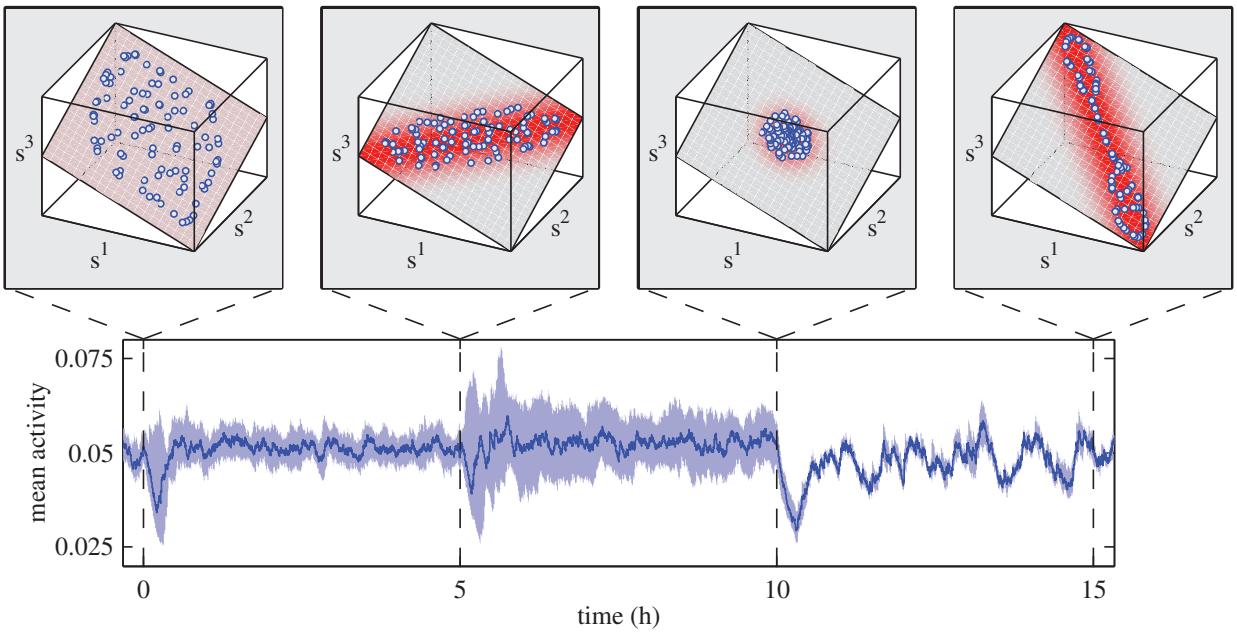


Figure 3.8.: The initial uniform distribution from which stimuli are sampled is repeatedly changed at time steps $t = 0$, $t = 5$, and $t = 10$. The insets depict the distribution of the developed receptive fields overlaid to the previously applied sampling distribution. The bottom panel shows how the average activity levels of the neurons develop over time. Therefore, the median as well as the upper and lower quartiles of the activity level distribution are plotted.

try to compensate for these changes in order to maintain stable activity patterns. More precisely, the less active neurons increase their sensitivity such that they become active for other stimuli as well. In other words, they expand their receptive fields and step in competition for responsibility in representing those other stimuli. The subsequent competition between neurons lets the receptive fields shrink again such that the neurons become selectively responsive to the new stimuli. As a result of this adaptation process, the average activity level of all neurons approaches the target level again. The inset at time $t = 5$ shows the distribution of the new receptive fields which nicely resembles the applied input pattern distribution.

In the following, we apply two more changes to the input pattern distribution: The first change occurs at time $t = 5$, where we start to sample input patterns according to $s^1 \sim \mathcal{N}(0, 0.04)$ and $s^2 \sim \mathcal{N}(0, 0.04)$. The second change is applied at time $t = 10$, from where on input patterns are sampled according to $(s^2 - s^1) \sim \mathcal{N}(0, 0.09)$. Both disturbances are visible as changed activity levels of the neurons. The homeostatic processes consequently force the neurons to reorganize their receptive fields. The insets at time $t = 10$ and $t = 15$ show that these reorganizations develop receptive fields whose distributions resemble the applied input pattern distributions.

These results illustrate the ability of the network not only to cope with changes in stimuli strength, but also to adapt to changes in the stimuli distribution. The latter is particularly interesting for modelling developmental systems, since learning-induced changes in the connectivity of the network can significantly alter the input pattern distribution.

Influence of the Target Activity Level Parameter

In the following, we evaluate the influence of different parameter settings. Due to the incorporation of homeostatic processes, the number of free parameters that have to be controlled reduces to the time constants as well as the target activity of individual neurons. We already discussed the time scale at which homeostatic processes have to operate. More precisely, these processes have to be fast enough to compensate for long-lasting activity changes induced by Hebbian learning. However, they additionally have to be slow enough in order not to destroy the moment to moment fluctuations which carry the input signal information. What remains is an investigation of the target activity parameter.

As already shown, the target firing rate determines the average activity levels of individual neurons. We consequently hypothesized that the target firing rate influences the size of the developed receptive fields. In this case, the parameter would further affect the overlap between the receptive fields of individual neurons and therewith the sparsity of the developed representation. To validate this hypothesis we performed multiple simulations using different target activity levels. The developed receptive fields for two of these simulations are exemplarily shown in Fig. 3.9. In (a) we see that at a target activity level of $\widehat{A} = 0.075$ neurons specialize to specific combinations of the three stimuli. In (b) we see that an increase in the target firing rate to $\widehat{A} = 0.15$ yields significantly larger receptive fields. To reach the target activity level individual neurons specialize to single input modalities. This is shown by the horizontal (s^1), vertical (s^2), or diagonal (s^3) response patterns of the neurons. We further calculated the overlap between the receptive fields for each parameter setting. The corresponding result is plotted in Fig. 3.10. Here, we observe a steady increase in the overlap when the target activity is increased.

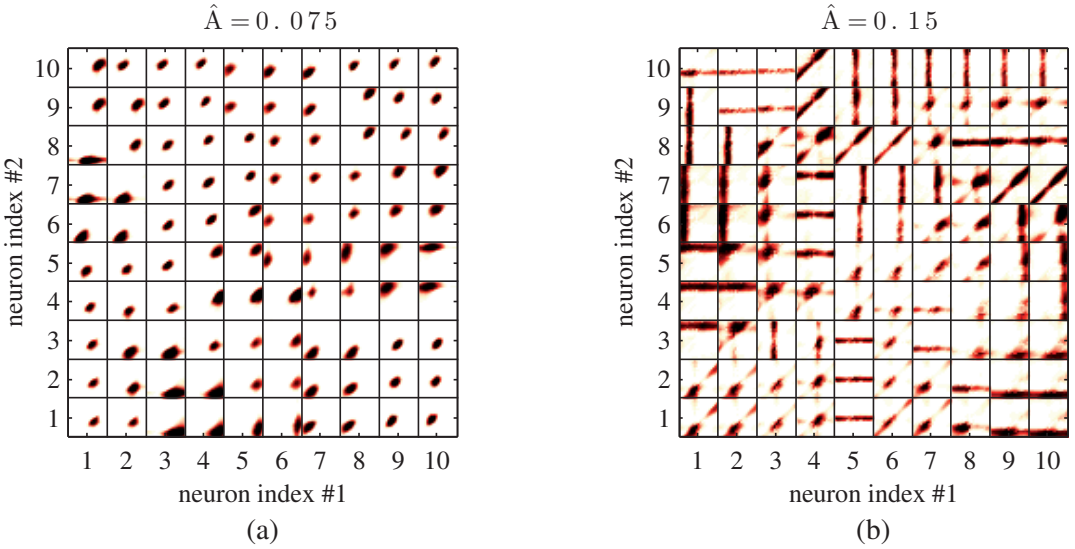


Figure 3.9.: The receptive fields of all excitatory units are shown for two simulation runs: using a target activity of (a) $\widehat{A} = 0.075$ and (b) $\widehat{A} = 0.15$. Here, each subimage corresponds to the response pattern of a particular neuron to different combinations of s^1 (x-coordinate) and s^2 (y-coordinate).

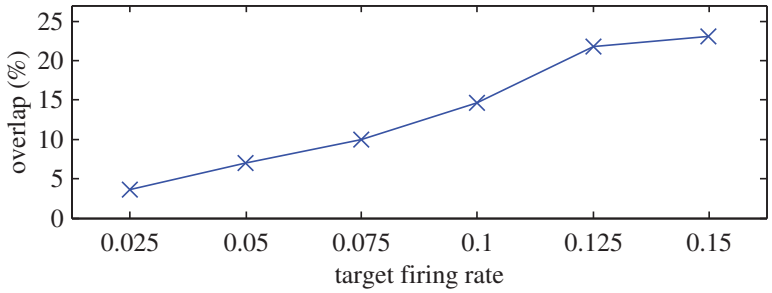
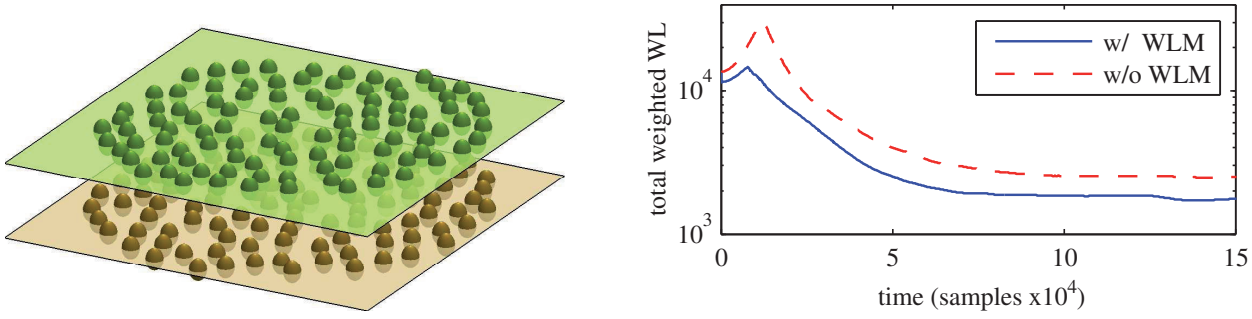


Figure 3.10.: The influence of the target firing rate parameter on the overlap between the developed receptive fields is depicted.

Topology Preservation

We next discuss the additional incorporation of the wiring length minimization (WLM) process. As described in section 3.2.4 the process adapts the neuron positions such that neurons with strong lateral connections become adjacent to each other. Since a lateral connection between model units only features a large synaptic weight when the neurons' receptive fields are similar to each other, WLM should facilitate the development of topology preserving mappings. When using WLM in the multi-modal association learning experiment, we finally obtain the spatial neuron layout depicted in Fig. 3.11 (a).

To estimate the effect of WLM on the topology preserving properties of the developed mapping, we compared the results of two simulation runs: one using WLM and one not using WLM. Here, we first investigate whether our interpretation of a neural field to be an elastic net (with lateral connections exerting attraction forces on units) is a suitable choice for minimizing wiring length. To do so, we calculated the total weighted wiring length (Eq. (3.17)) for both simulation runs. To compensate for different spatial scales we further normalized the distances between units by their mean. The evolution of the resulting measure is depicted in Fig. 3.11 (b), where the locarithmic scaling of the y-axis should be noted. The total wiring length of the simulation without WLM increases at the



(a) (b)

Figure 3.11.: (a) shows the final layout of the model units as obtained when using the wiring length minimization process during neural field formation. (b) depicts the evolution of the normalized total weighted wiring length for a simulation using WLM as well as a simulation not using WLM.

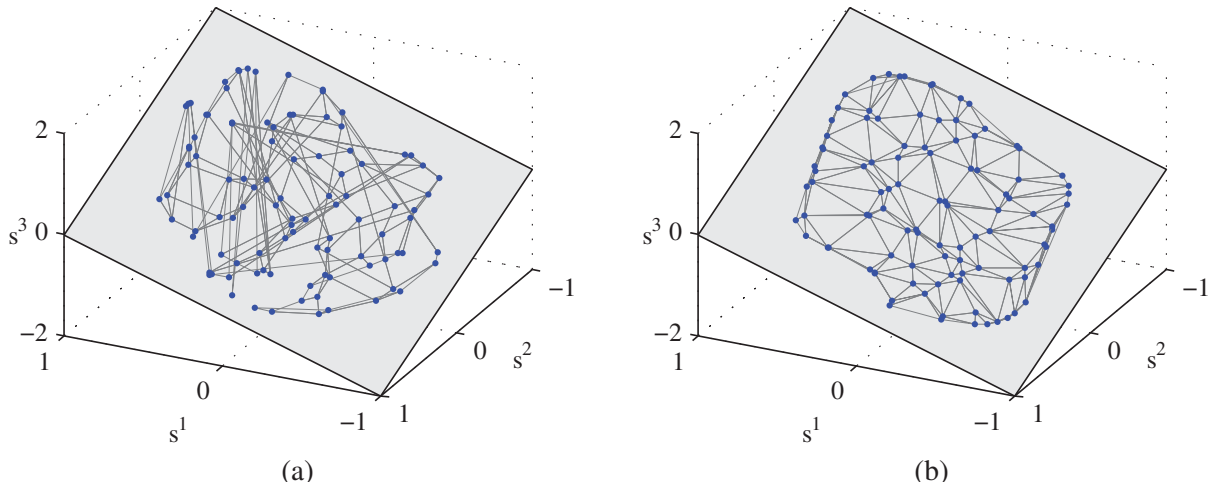


Figure 3.12.: The center positions of the developed receptive fields are plotted. Connections are drawn between those receptive fields, whose neurons are adjacently positioned in the output plane. (a) shows the result of a simulation where WLM has not been used, whereas the result depicted in (b) has been obtained using WLM.

beginning. This can be attributed to an initial rough adjustment of the lateral connection weights. A competition between the model units subsequently induces a "die off" of many synapses, which let the total weighted wiring length decrease over time. When incorporating WLM we observe a similar trend, except that the initial increase in total wiring length vanishes. Most importantly, however, our implementation of WLM results in a smaller total weighted wiring length.

Given the ability of our model to reduce the weighted wiring length between units, we now demonstrate that WLM is suitable for improving topology preservation. Therefore, we compare the mappings which have been developed by the two simulation runs. To do so, we first calculate the neighborhood relations between the excitatory units using Delaunay triangulation of their positions after training. We additionally calculate the positions of the centers of the developed receptive fields. The resulting receptive field positions are plotted in Fig. 3.12, where we overlaid connections between them according to the calculated neighborhood relations. For a topology preserving mapping this would result in a plot where neighboring receptive fields are connected (due to the adjacent positions of their corresponding neurons). As shown in Fig. 3.12 (b), this is the case for the neural field which has been trained using WLM. In contrast, not using WLM results in significant topological defects (see Fig. 3.12 (a)).

These qualitative results can be confirmed by a quantitative analysis using the topographic function (Villmann et al., 1997). This widely used measure characterizes the topology preservation of mappings by analyzing the degree of topological defects on varying scales: from local to global ones. The results are plotted in Fig. 3.13. There, the normalized rank k determines the effective neighborhood range, i.e. small $|k|$ correspond to a local neighborhood, whereas large $|k|$ correspond to a global one. The results show that WLM decreases the number of topological defects on both a local scale and particularly on a global scale.

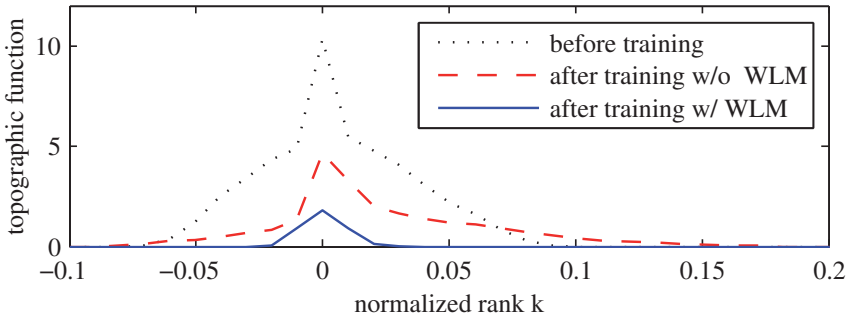


Figure 3.13.: The topographic function is plotted for neural fields which have been trained with WLM or without WLM.

3.3.2. Development of Phoneme Concepts

We finally apply our network model in the domain of speech processing. Therefore, we present results of simulations where the neural field has been trained using continuous speech input. The model consequently should develop a mapping where individual neurons specialize to specific sounds, i.e. phonemes. By incorporating the WLM process, the mapping should further maintain topology, i.e. similarly sounding phonemes should be mapped onto neighboring neurons.

Since human speech perception relies to a large extent on vocal tract resonance frequencies and their variation in time (Furui, 1986), we decided to use formant frequencies as input to our model. Thereby, formants refer to the energy concentrations in the spectro-temporal domain (which are correlates of the underlying vocal tract resonance frequencies). It has been shown that formant trajectories can be robustly extracted from continuous speech (Gläser et al., 2010a; Heckmann et al., 2008). In the present experiment, however, we use the hand-labeled trajectories provided by the VTRFormant database (Deng et al., 2006). As a subset of the widely used TIMIT corpus, this database comprises a total of 516 utterances from which we used all 322 utterances spoken by male speakers.

We use a population code of 128 neurons to represent the formant frequencies at each time frame. Thereby, the response patterns of the input neurons correspond to the transfer functions of a 128-channel Patterson-Holdsworth auditory filter bank (Patterson et al., 1992). This filter bank is based on neurophysiological findings on the human auditory system and models the peripheral processing as carried out by the cochlea, where sound is transformed into spatio-temporal response patterns on the auditory nerve. In our setup the filter bank is composed of Gammatone filters whose logarithmically arranged center frequencies cover the range from 80 Hz to 8 kHz. An exemplarily selected speech utterance is shown in Fig. 3.14. There, (a) depicts the time-domain signal, (b) the corresponding formant trajectories coded by the population of input neurons, and (c) an example input pattern at the specific time frame marked in (b). We constructed the input samples using a sampling rate of 1 kHz. The remaining setup of our network model equals the one described in Section 3.3.1.

To illustrate the benefit of using WLM we once again trained two neural fields on the speech data. The positions of neurons in the first field have been fixed to a 10x10 grid, whereas

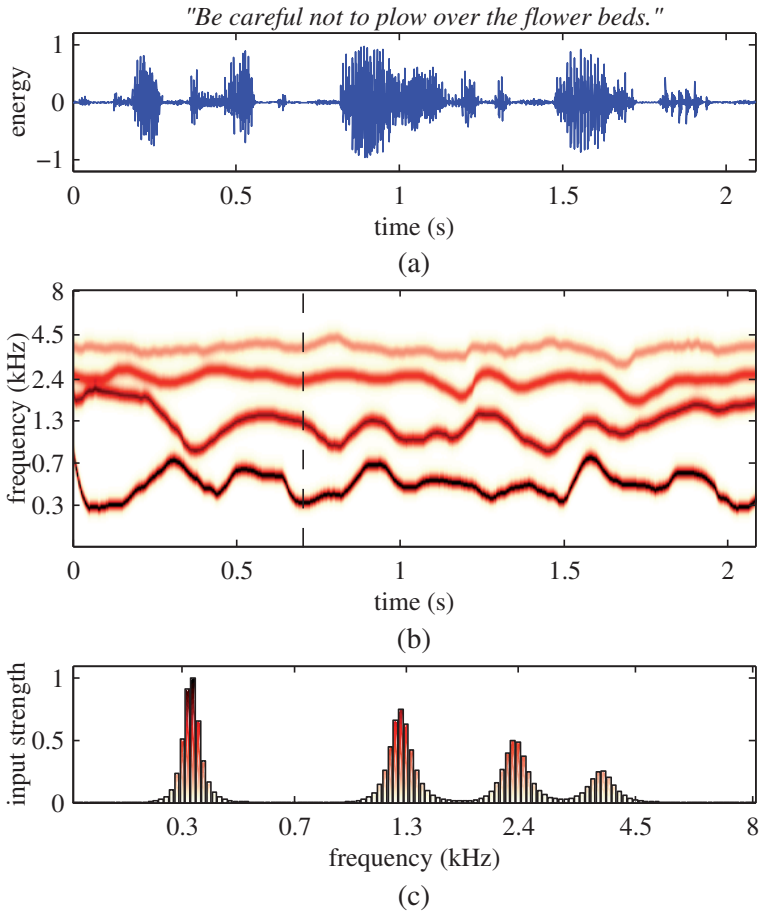


Figure 3.14.: In (a) the time-domain signal for a speech utterance is shown. The corresponding formant trajectories are depicted in (b), whereas (c) shows the population code of the formant frequencies at a specific time frame.

neurons of the second field could change their positions using WLM. After training we analyzed the response of the fields to different input stimuli. More precisely, we calculated the mean formant frequencies for each phoneme transcribed in the VTRFormant database and recorded the neuron responses to the corresponding input patterns. Finally, this allowed us to label a neuron with the symbol of the phoneme which evokes the largest response of the neuron. In Fig. 3.15 we plot the final layout of the excitatory neurons as well as their labels. For ease of interpretation we restricted the labels in the plot to vowels and semivowels. The results depicted in (b) illustrate that the incorporation of WLM produces a topology preserving mapping where neurons with similar labels cluster together. Furthermore, the labels are distributed over the map such that similar sounding phonemes are close to each other. In contrast, (a) shows that a training of the field without using WLM produces a map, where symbols of similar sounding phonemes are widely distributed. This difference also becomes evident in the response patterns to single phonemes, e.g. illustrated in Fig. 3.15 (c) and (d) where we plot the field activity following the presentation of an "i". More precisely, the field constructed without using WLM produces multiple loci of activity, whereas the use of WLM forces the formation of a map exhibiting single activity bubbles.

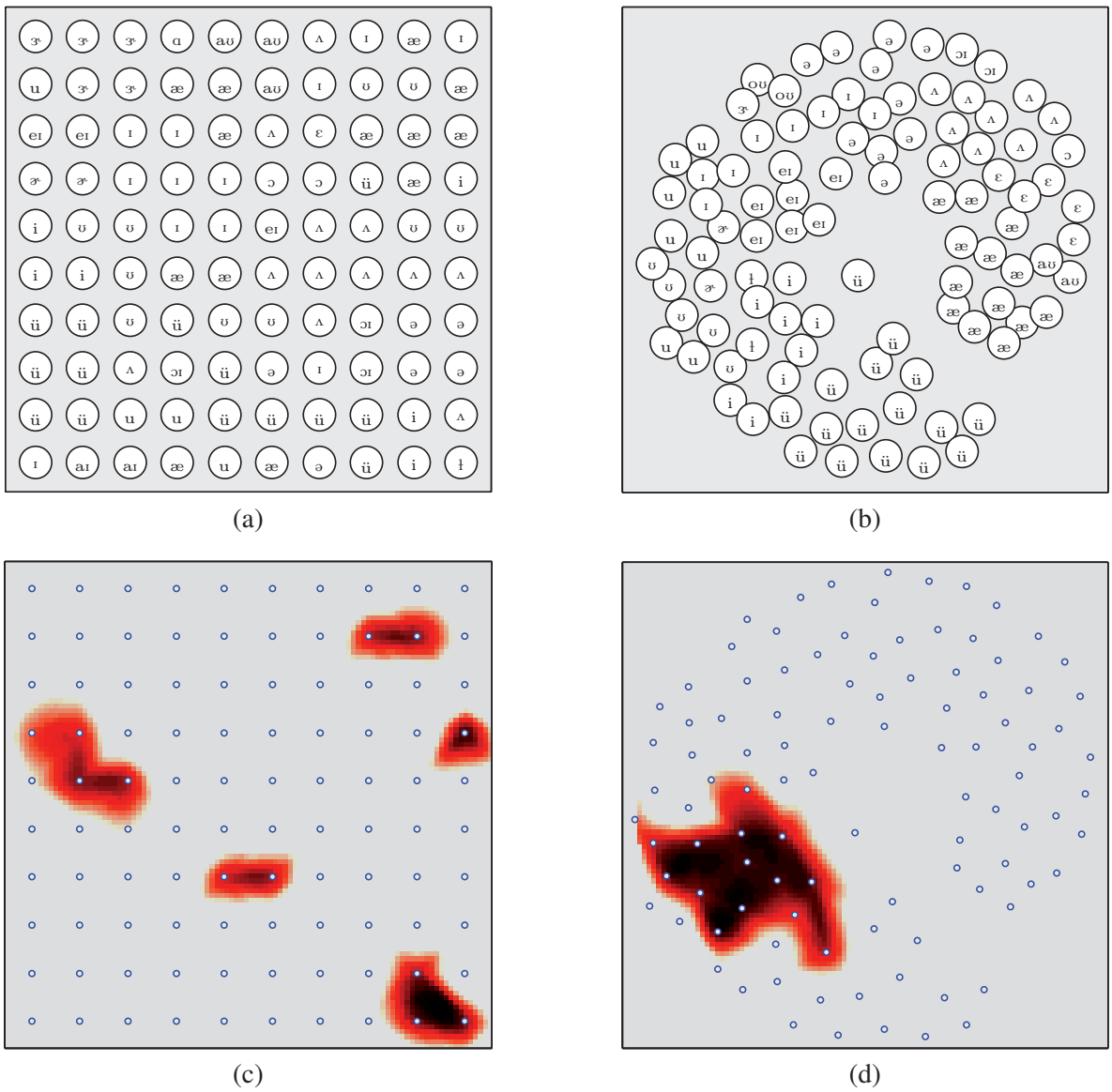


Figure 3.15.: The positions of the excitatory units, when the field has been trained (a) without WLM and (b) with WLM. Phonemes (vowels and semivowels), to which the neurons exhibit the largest response, have been used as labels for the neurons. Phoneme labels are from the IPA phonetic alphabet. Plots (c) and (d) depict the field activities as evoked by the phoneme "i".

The development of topology preserving mappings is particularly advantageous for the processing of speech, since a continuously changing stimulus (e.g. formant trajectories) evokes a continuous activity trace in the field. This is illustrated in Fig. 3.16, where we plot the trace of neurons which exhibit the largest response to the word "money" [m-ʌ-n-i]. In (a) we see that a mapping with many topological defects does not produce a continuous activity trace, whereas the topology preserving mapping in (b) does. We further added jitter to the plot. This allows us to estimate the time course at which the activity bubble moves from one position to another. From the plot it becomes evident that the activity bubble remains at positions, which correspond to the phoneme cores, for

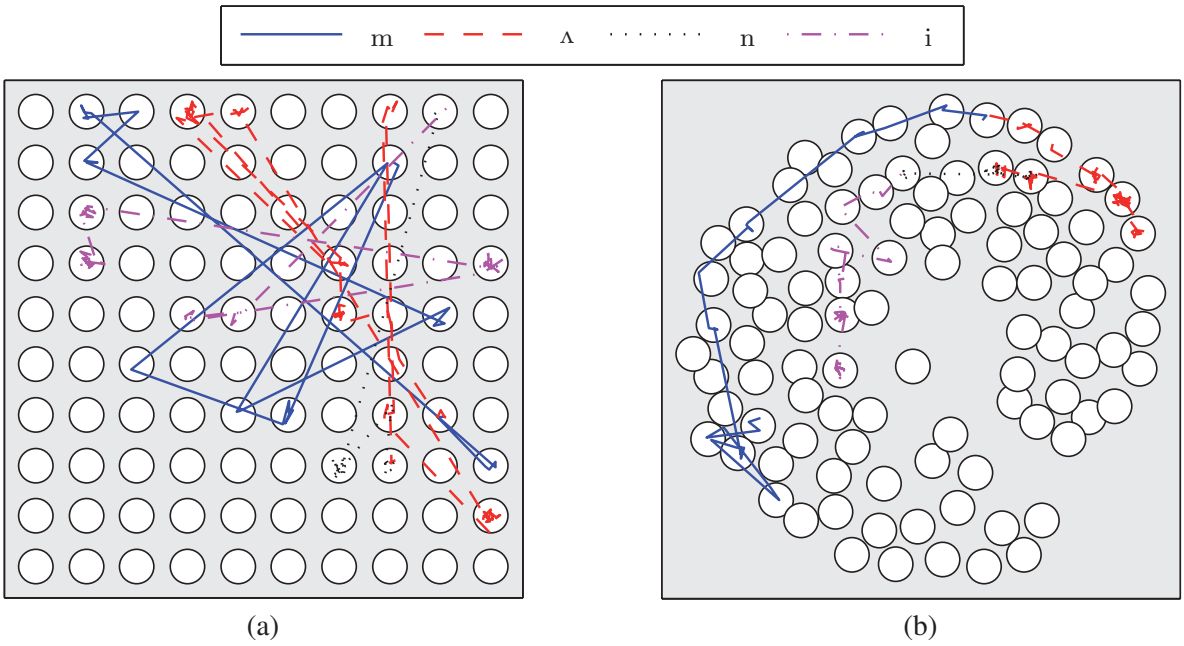


Figure 3.16.: The trace of neurons which exhibit the largest response to the word "money" [m-ʌ-n-i]. (a) shows a non-continuous trace for the map developed without WLM, whereas (b) depicts a continuous activity trace for the map developed with WLM.

a relatively long period of time. In contrast, a fast movement of the activity bubble is observed for transitions between phonemes. A higher-level processing of speech, e.g. a phonetic transcription (Kohonen, 1988) or a speech synthesis (Vaz et al., 2009), could be based on such traces of activity.

3.4. Application to Word Learning

The evaluation in the previous section assessed the computational characteristics of the homeostatic DNF. This means it allowed us to investigate whether and how the model accomplishes the goals of unsupervised learning, e.g. pattern discovery, topology preservation, and adaptivity. What remains is to demonstrate the model's application to word learning. In this section, this is done in the color domain. More precisely, it is shown how the model can be used to develop color concepts in a data-driven way and to subsequently link corresponding word labels to them.

This word learning scenario has been chosen for two reasons. Firstly, color obviously is one of the domains, where concepts are acquired in an unsupervised way. Children observe colors from the very beginning of their life and are able to distinguish between them far earlier than the onset of speech. They hence develop color representations solely based on visual observations without the need of a supervision by words. Secondly, however, words can significantly alter these color representations afterwards. Whereas the initially developed color categories seem to be universal for different cultures (Berlin and Kay, 1991), language subsequently affects the ability to discriminate between them, i.e.

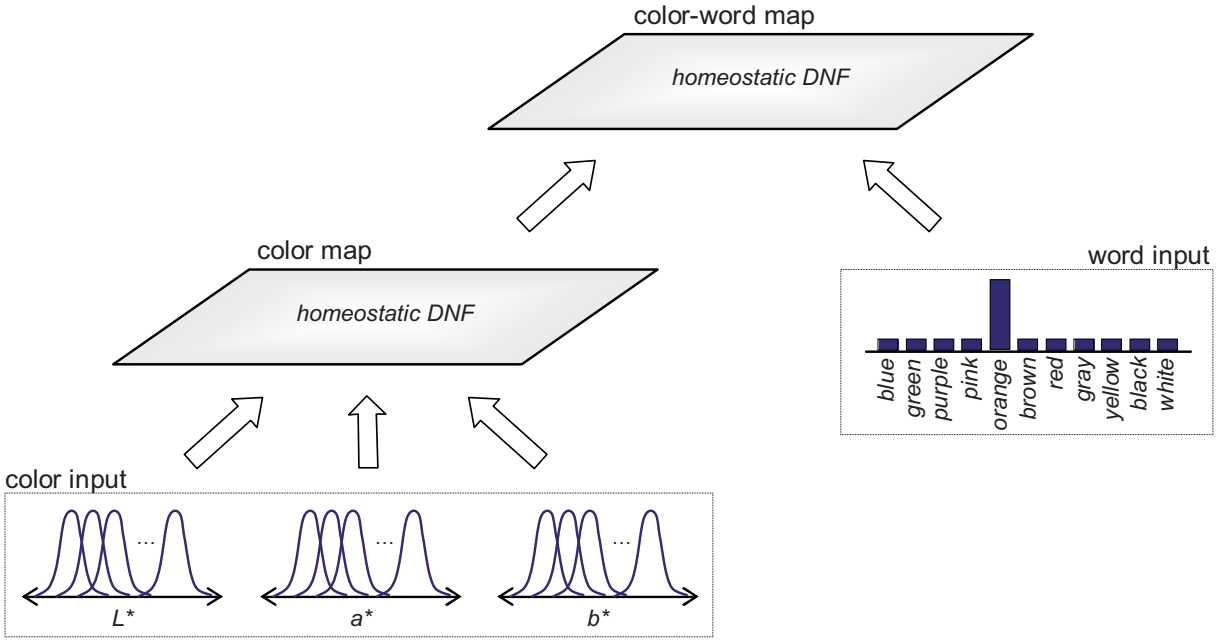


Figure 3.17.: The architecture of the system used for color concept development.

people with different color vocabularies perceive colors differently (Roberson et al., 2000; Davidoff, 2001; Kay and Regier, 2006). In the following, we show that the homeostatic DNF model can resemble these results.

Fig. 3.17 depicts the system architecture that is used within the scenario. The system is composed of two successive DNF layers. Thereby, the first DNF is meant to provide a *color map*. Its development is solely driven by visual observations, i.e. color values, based on which the DNF should develop an appropriate color representation. Input to the second DNF is provided by the first DNF as well as by corresponding word labels. The second layer hence constitutes a *color-word map* which integrates purely visual input and purely auditory input, thereby establishing color categories. Both DNF layers only learn bottom-up, i.e. they are not recurrently coupled.

Color input to the first layer is represented in the *CIE L*a*b** color space. This is a 3-dimensional space, where L* denotes the lightness channel, a* the red-green channel, and b* the yellow-blue channel. We choose this color space, since it reasonably approximates human perception. More precisely, the Euclidean distance between two colors in CIE L*a*b* correlates with the perceived difference between them (Tkalcic and Tasic, 2003). Each of the three color channels is coded by a population of 21 input neurons with equidistant Gaussian-shaped receptive fields yielding a 63-dimensional input in total.

The words, that are supplied to the second DNF, provide the labels for the different color inputs, i.e. they constitute color names. To construct appropriate word labels we relied on the data of a color naming study conducted by Berlin and Kay (1991)[1]. In the study, 20 participants – each of them speaking a different language – were asked to label the Munsell chips depicted in Fig. 3.18 (a). The English speaking subject labeled a subset of

[1]The database can be downloaded from the World Color Survey Data Archives at `http://www.icsi.berkeley.edu/wcs/data.html` (29.08.2011)

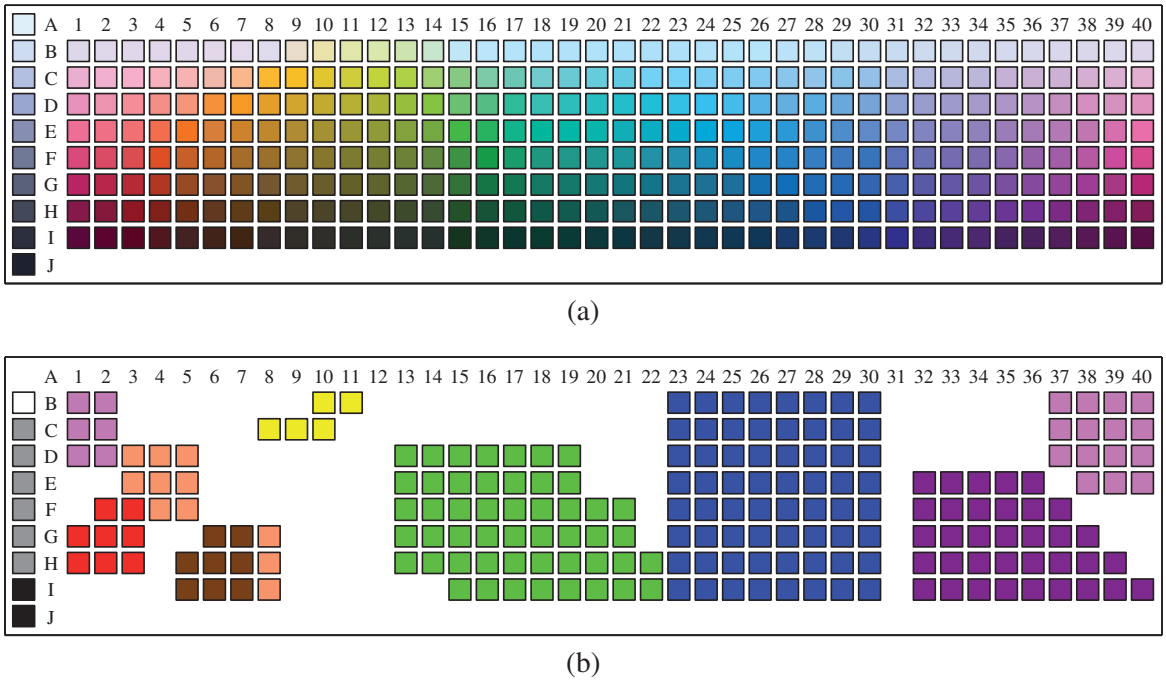


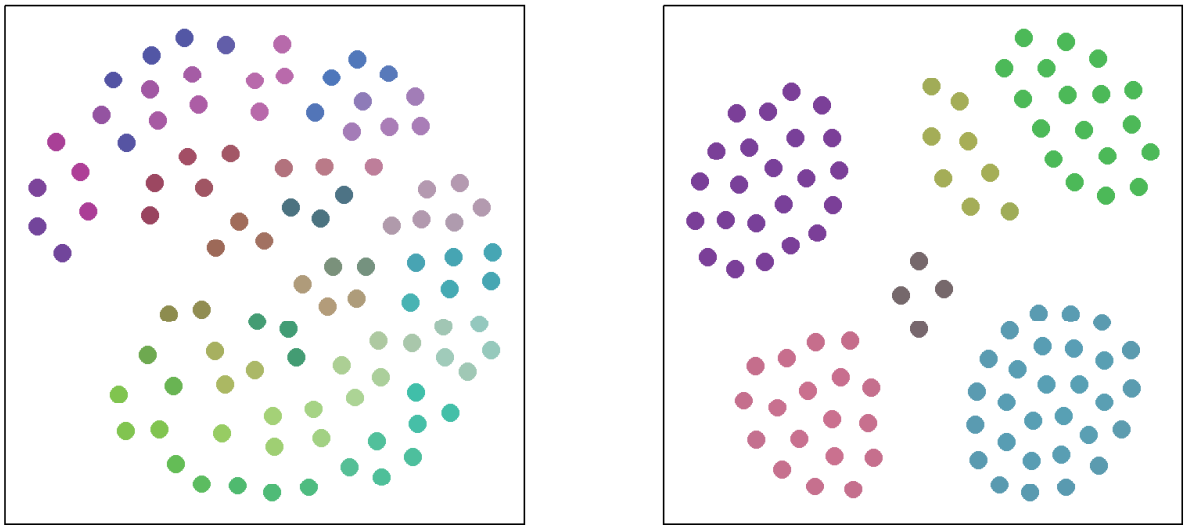
Figure 3.18.: The data of the color naming study of Berlin and Kay (1991): (a) shows the 330 Munsell chips that have been presented to the participants. (b) depicts how the English speaking person labeled 211 of them using 11 different color terms.

the chips as depicted in (b), thereby making use of 11 basic English color terms. The terms as well as their frequency of occurrence are summarized in Table 3.1. We used the labeled Munsell chips as prototypical color-word mappings. This means, for an arbitrary color input to the first DNF, we calculated the most similar Munsell chip that has been labeled by the subject. Finally, the corresponding word label is used as input to the second DNF. Therefore, the word input to the second layer is composed of 11 neurons, only one of them being active for each color input, respectively.

The two DNFs are composed of 100 excitatory and 100 inhibitory neurons each. We used a target firing rate $\widehat{A}$ of 0.1 and further applied wiring length minimization during training. The remaining network parameters were chosen as described in Section 3.3. In the simulation, we randomly sampled color input values and calculated the corresponding word labels. The resulting color-word pairs were sequentially presented to the system. For each input we allowed the networks to propagate activity for 50 iterations before the next pair has been applied. Total training time included the presentation of 20000

blue	green	purple	pink	orange	brown	red	gray	yellow	black	white
64	50	35	21	11	8	8	6	5	2	1

Table 3.1.: The 11 English color terms used in the study of Berlin and Kay (1991) as well as their frequency of occurrence.



(a) (b)

Figure 3.19.: The spatial layout of the two DNFs after training. Neuron colors depict the input colors the individual neurons are most responsive to. Whereas (a) shows that the *color map* develops a continuous color representations, (b) shows that the *color-word map* forms color categories.

input pairs (even though similar results could be obtained using approx. 3000 samples). After training, the network parameters were fixed and the developed representation were investigated.

The evaluation is based on a test set that has been constructed using a fine-grained equidistant sampling of the input color space. While presenting each test input to the network, the responses of all neurons were recorded. This yields the receptive fields of the different map neurons. We further calculated the centers of the individual receptive fields by which we obtained the codebook vectors of the units, i.e. the colors the neurons are most responsive to. In Fig. 3.19 we plot the spatial layout as well as the individual codebook vectors (color-coded) for the *color map* and the *color-word map*, respectively. From the plot in (a) it can be seen that the color inputs drive the first layer DNF to develop a map in which colors are represented as a continuum. More precisely, any input color is covered by the neurons and, moreover, a topographic mapping is achieved. As a result, the system is able to recognize all colors and can further distinguish between them. The more distant the respective color representations are, the more different the colors are perceived. The contrary is true for the color-word map in (b). The inclusion of word labels seems to drive the neurons to cluster functionally. Interestingly, this clustering cannot only be observed in terms of the neurons' response characteristic but also with respect to their spatial arrangement. In summary, the color-word map forms categories in which one and the same codebook vector is shared among the category neurons, respectively. This means that focal (within-category) colors can be more stably judged than out-of-category colors (Mervis et al., 1975). The transition from an initially continuous towards a categorical perception of color is known to be present in infant development (Roberson et al., 2004). The model thus produces reasonable results.

An investigation of the input weight values of the second layer DNF revealed that the *color-word map* formed categories for *blue*, *green*, *purple*, *pink*, *yellow*, and *gray*. Other word labels were not represented by the map. The frequency of label occurrence obviously seems to be one important aspect in this respect. The 4 most frequent labels formed categories within the map (cf. Table 3.1). However, this argumentation does not hold for the cases of *yellow* and *gray*, since other words, e.g. *orange*, are more frequent than them. Our explanation for the observed behavior is that the map tries to balance between representing frequent words and appropriately covering the color space. In other words, the color *orange* is very similar to the already represented color *pink*. The contrary is true for *yellow*, which is largely different from any of the other colors. Therefore, the map prioritizes an appropriate vector quantization of the color space over the frequency of word occurrence in this case. The same is true for *gray*.

Finally, an analysis of the neuron responses showed a significant correlation for the neurons that belong to the same category. More precisely, all neurons of a category are most active if the focal (codebook) color is input to the system. In contrast, the responses of the neurons decline when the stimulus deviates from the focal color. Neuron activity hence codes how well an observed color matches the category prototype. However, this also means that the ability to discriminate between colors of the same category is impaired. This is because two colors can be different (e.g. *light blue* and *dark blue*), but have the same distance to the prototype (e.g. *blue*) and hence result in similar activities. The contrary is true for colors at category boundaries. There, discrimination is enhanced, since different neuron populations represent the different colors. This behavior – also known as *perceptual magnet effect* – is one of the hallmarks of categorical color perception in humans (Davies et al., 2003).

In summary, the results showed that the homeostatic DNF appropriately models color concept development. Firstly, the network achieves a self-organization of color representations solely based on visual stimuli. Secondly, the network resembles the effect that word labels have on these previously developed representations. If one considers word labels to emerge later than visual stimuli, the system hence results in a transition from an initially continuous to a finally categorical perception of color. Thereby, the properties of the developed color categories are in-line with findings about human color perception. Additional experiments could strengthen this statement, but are left for future research.

3.5. Discussion

Unsupervised concept formation strives for a data-driven self-organization of categorical representation. This includes an extraction of reoccurring input patterns and their embedding into topographic maps, but also the maintenance of stable representations in face of changing external environments. The human brain achieves these goals marvelously well. Therefore, it is no wonder that multiple computational models exist, which take inspiration from brain-like processing principles. *Kohonen Maps* and *Dynamic Neural Fields (DNFs)* are such popular techniques. They already have been applied in a variety of tasks and were of primary interest in the present work, too. Whereas Kohonen Maps captivate by their algorithmic simplicity, easy use, and yet powerful results, DNFs are of

interest due to their computationally attractive feature of dynamic activity propagation. The dynamical aspect, however, is also a drawback of DNFs, insofar as the networks are sensitive to parameter settings and hence difficult to use. This problem is even more severe when network parameters are altered via learning.

In this chapter we presented a computational model that aims at bridging the gap between the two techniques. More precisely, the recurrent network is able to integrate successive inputs in a dynamical fashion, propagates activity via extensive lateral connection and hence maintains the attractiveness of DNFs. The network is further able to learn and adapt a topographic representation which are aspects that can hardly be achieved using conventional DNFs. The incorporation of homeostatic principles is particularly important in this respect. These mechanisms self-regulate the network activity, maintain a stable operation mode, and thus circumvent the parameter sensitivity that hinders standard DNFs. As a result, the network is as easy to use as Kohonen Maps are, while relying on dynamical processing as DNFs do.

We thoroughly evaluated the model in a series of experiments. Firstly, artificially generated data has been used to address multi-modal association learning in the domain of reference frame transformation. Secondly, we investigated the development of phoneme representations following continuous speech input. The results demonstrated that the network self-organizes without any external supervision, develops appropriate representations of the input, and even adapts to sudden changes in the strength or distributions of input patterns. We could further show that wiring length minimization significantly enhances the quality of the developed mappings with respect to topology preservation.

We finally showed how the network is able to acquire the meaning of words. Color names served as an example in this respect. In detail, the model was used to first develop a color representation that is independent of word labels. Subsequently, this representation has been linked to the words via an additional processing layer. We could show that the formation of color categories by the network resembles aspects that are known from the development of color categories in children.

4

Supervised Word Meaning Acquisition

Language shapes the way we think, and determines what we can think about.

Benjamin Lee Whorf (1897-1941)

Children are astonishing word learners. Particularly the formation of concepts for newly heard words does not seem to constitute a problem for them. Children rather seem to possess dedicated learning mechanisms that allow them to efficiently master this task. This becomes evident by a number of observations. Firstly, children rapidly get a glimpse on the meaning of a novel word. Secondly, just a few occurrences of a word enable children to generalize its meaning. And, finally, children quickly integrate new words into their lexicons and successfully use them in their own discourses. Artificial learning systems strive for similar child-like abilities but ultimately fail to achieve them. Hence, the question what kind of learning principles children apply and whether computational systems can benefit from similar mechanisms naturally arises.

This chapter aims at providing answers to these questions. It first reviews supervised concept formation from the perspective of developmental psychology and further establishes a link with neurobiological learning theories. The discovered principles are used to infer the computational requirements for child-like word learning in artificial systems. In the following, a novel computational model that internalizes these principles is presented in detail. The benefit of a bio-inspired system over conventional approaches is assessed via thorough evaluations of the model in a wide range of problems. An application of the model in a simulated word learning scenario finally examines whether the artificial system exhibits learning dynamics similar to those observed in children.

4.1. Word Learning Processes in Infants

An efficient learning of word meanings entails that contradictory needs are satisfied simultaneously. On the one hand, a system should acquire the meaning of new words from few training exemplars. This is necessary, since an exhaustive teaching by a tutor is long-lasting, undesirable, and often even not possible. On the other hand, the gathered word knowledge should concentrate on the 'core' meaning, i.e. it should be independent of the specific context in which a word has been learned. This presupposes a process of abstraction that exploits the statistical evidence of multiple training exemplars. Child-like word learning can serve as a model for both aspects. The identification of the learning principles applied by children is of particular importance in this respect.

4.1.1. Fast & Slow Mapping

Psychologists assess children's word learning capabilities by confronting them with new words during natural social interactions. Careful designs of such experiments allow the identification of factors that may have an influence on children's word learning success. In a seminal study, Carey and Bartlett (1978) conducted an experiment to investigate the learning of a novel color term in 3- to 4-year-old children. More precisely, the children have been asked to *"Bring me the chromium tray, not the blue one, the chromium one"* while they walked toward a blue and an olive tray. Each tested child correctly inferred that *"chromium"* referred to the olive color. Surprisingly, even in a test performed six weeks after training many children still remembered the meaning of the word. Carey and Bartlett (1978) took this result as evidence for a rapid word learning mechanism and called this process *fast mapping*. Following the *chromium study*, fast mapping was in the focus of many experiments. The results often confirmed the findings of Carey (1978) such that nowadays fast mapping constitutes one of the hallmarks of word learning by children. At its core, it refers to the fact that just a few observations of a word (or even a single exposure to it) allow children to establish a word-meaning link; even though this link initially may not cover the complete word meaning. Fast mapping seems to rely on a general learning principle rather than a word-specific processing. It is not limited to labels, but also applies to the learning of facts. Markson and Bloom (1997) either used an object name (*"koba"*) or a fact (*"the thing my uncle gave me"*) to refer to an object. The results demonstrated that children fast mapped the object to the label as well as the fact. Moreover, fast mapping even seems to rely on processing circuits that are not specific to humans, since bonobos (Lyn and Savage-Rumbaugh, 2000) as well as dogs (Kaminski et al., 2004) have been reported to show similar fast mapping capabilities.

Because most early word learning studies concentrated on fast mapping, it has long been neglected that word learning does not only involve the acquisition of an initial word meaning. Questions on how such initially partial meanings are completed and finally retained over longer periods of time just recently entered the focus of infancy research. Horst and Samuelson (2008) reported that 2-year-olds robustly fast mapped labels to objects but showed deteriorated performance when retention was tested after a five minute delay. The authors suggested that a competition between multiple novel object-label associations may underlie this observation. In other words, in order to retain

an initial word-meaning mapping, it either has to gain support via explicit biasing (e.g. by highlighting the correct referent of a word by the caregiver) or by repeatedly observing the association in subsequent interactions. As this is an enduring process that gradually extends and strengthens an association, the mechanism has been termed *extended mapping* or *slow mapping*. According to Carey (2010) two factors primary influence this extended mapping process. Firstly, the need for the creation of new semantic primitives and, secondly, the size of the hypothesis space. The former aspect already has been mentioned in Section 2.4, where the importance of creating new representations (both in terms of conceptual categories as well as features associated with them) has been emphasized. The latter aspect refers to the fact that the learning of words with a concrete meaning requires less cognitive abilities than that of more abstract words. A good example in this respect is that nouns are typically learned earlier (or faster) than verbs. Here, the proposal is that nouns refer to specific entities in the environment, whereas verbs refer to actions that usually possess a high variability (e.g. in terms of how they are executed, who is doing something, or which objects are involved). Finding commonalities among the different situations, in which a word has been observed, consequently is more difficult to achieve for verbs than for nouns (Gentner, 2006; McDonough et al., 2011).

In summary, children's word meaning acquisition is characterized by two stages. Firstly, *fast mapping* creates an initial link between a word label and aspects of a scene in which the word occurred. This association is rapidly memorized in one-shot, but typically constitutes just a partial word meaning that is incomplete and fragile. Fast mapping results in a word meaning that is context-dependent, i.e. very much bound to the situation in which the word occurred. Multiple fast mapped association, however, serve as a basis for abstraction during a subsequent *slow mapping* process. A child can compare the different associations and search for commonalities among the referents of a word. This is an enduring process by which a word's meaning is gradually decontextualized. Hence, the acquired concept finally describes the essential meaning of a word with all context-dependent semantic primitives being removed.

4.1.2. Complementary Learning Systems Theory

Unveiling the biological circuits underlying fast and slow mapping may enable us to construct artificial systems that use learning principles similar to those employed by children. Unfortunately, establishing such an interdisciplinary link has rarely been tried. In the following, it is argued that *Complementary Learning Systems (CLS)* theory (McClelland et al., 1995) provides one plausible explanation on the neurobiological basis of fast and slow mapping. Therefore, we first review CLS theory before a number of findings from both developmental psychology and neurobiology are presented to underpin this proposal.

The CLS account constitutes a general theory on learning in the brain. Thereby, a key role is given to the notion of complementary learning and memory systems. More precisely, these systems are the hippocampus on the one hand and neocortical areas on the other. According to McClelland et al. (1995) both systems possess unique computational characteristics that individually serve complementary purposes:

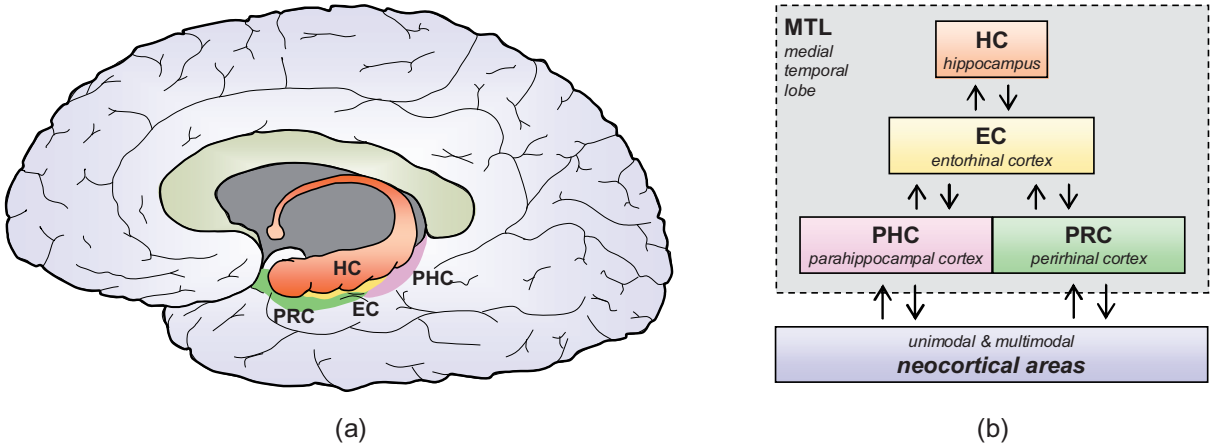


Figure 4.1.: The medial temporal lobe (MTL): (a) shows the locations of the different MTL structures, whereas (b) depicts their connectivity.

- The hippocampus is part of the allocortex. In conjunction with the entorhinal cortex (EC), the parahippocampal cortex (PHC), and the perirhinal cortex (PRC) it forms the medial temporal lobe (MTL). As illustrated in Fig. 4.1, many neocortical regions project to the hippocampus via the other MTL structures. The hippocampus consequently constitutes a region in which associations between arbitrary modalities can be formed (Squire, 1992). Not only its prominent location is what distinguishes the hippocampus from other brain areas, but also its capability of rapid learning. Declarative memory like facts, autobiographical events, or semantic associations can be acquired in one-shot (Bliss and Collingridge, 1993). According to McClelland et al. (1995) the basis for this unique ability are localized representations. They allow the hippocampus to memorize different items with minimal interference between them. This means that the problem of *catastrophic forgetting* – the overwriting of existing memories by the encoding of new ones – is circumvented.

- The neocortex comprises many different areas. Some of them are unimodal (e.g. the primary sensory cortices) whereas others process multimodal information (e.g. the prefrontal cortex). What they have in common are the distributed knowledge representations they employ, i.e. any information processed by neocortical areas reflects itself as an activity pattern that is distributed among a cell assembly. According to McClelland et al. (1995) these overlapping representations enable neocortical areas to efficiently represent knowledge. However, the increased memory efficiency comes at the cost of decreased learning speed, since a slow learning from an interleaved presentation of overlapping activity patterns is required to avoid catastrophic forgetting.

The CLS theory suggests that the human brain combines the two learning systems to achieve a rapid encoding as well as an efficient storage of memories. Thereby, learning comprises the three steps that are illustrated in Fig. 4.2. In (a) it is shown that new observations in form of distributed activity patterns in neocortical areas are first encoded in the hippocampus. Therefore, the hippocampus 'allocates' new memory items which act akin to 'pointers' towards these patterns. Secondly, (b) depicts the reactivation of such

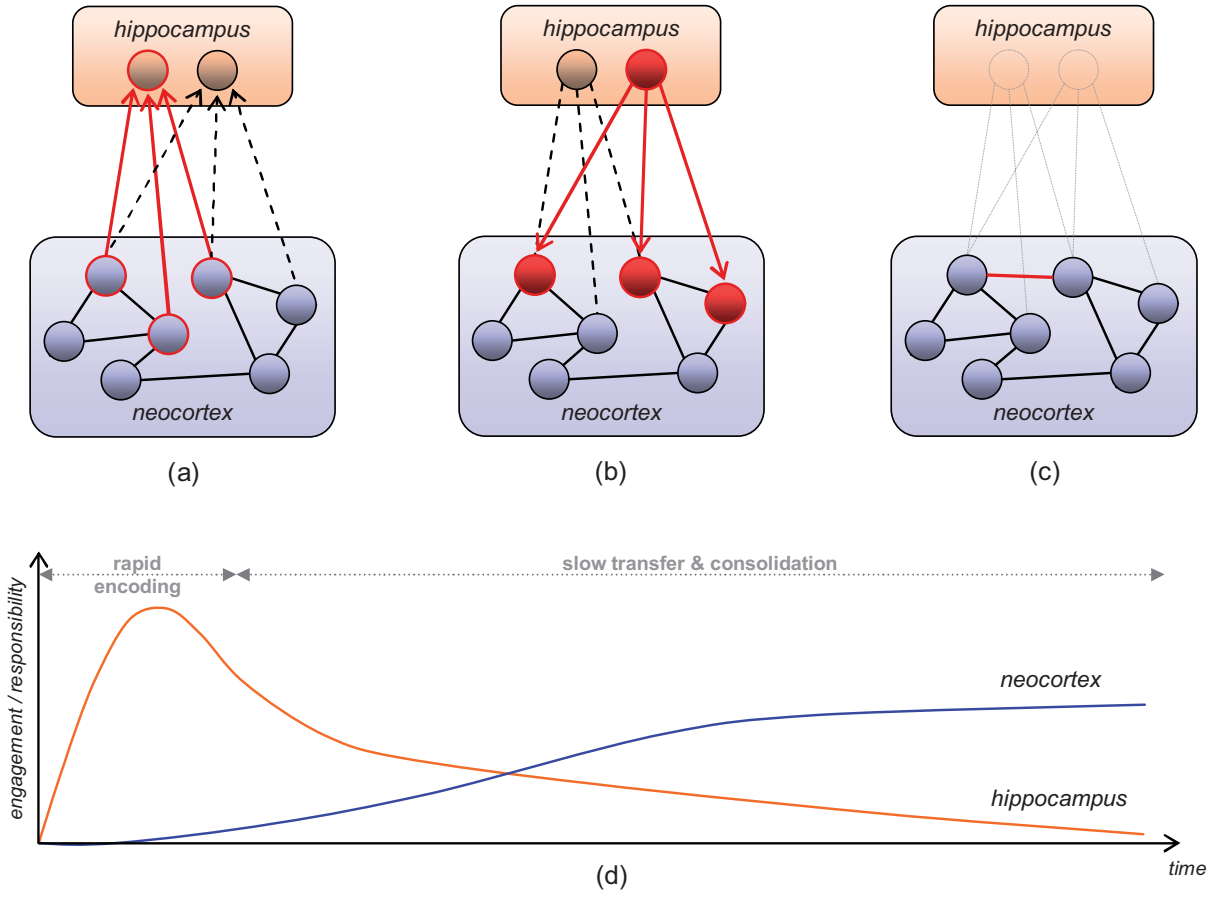


Figure 4.2.: Illustration of the biological mechanism underlying the acquisition of new memories and their consolidation over time. CLS theory comprises (a) the rapid encoding of activity patterns in the hippocampus, (b) their subsequent reactivation during sleep, and (c) the knowledge transfer to neocortical sites for long-term storage. This results in a gradual memory consolidation as illustrated in (d).

patterns. Due to the recurrent connectivity between the hippocampus and neocortical areas, the hippocampus can replay the different stored associations. This process occurs during sleep or rest (Ji and Wilson, 2007; Carr et al., 2011) and reactivates previously stored activity patterns in a quasi-parallel manner. As illustrated in (c), this finally allows neocortical areas to adapt their representations and to incorporate new knowledge via slow interleaved learning. Accordingly, the hippocampus' primary role is that of a short-term memory which can rapidly store new items. In contrast, neocortical areas store long-term memories whose acquisition is driven by the reactivation of patterns from short-term memory. This coupling between both systems results in a memory consolidation process that gradually transfers knowledge from the hippocampal system to neocortical sites (Frankland and Bontempi, 2005). As illustrated in Fig. 4.2 (d), hippocampal engagement during memory tasks is thus large for recently acquired knowledge, but it steadily decreases as knowledge transfer progresses and neocortical areas take over responsibility (Takashima et al., 2006).

CLS theory previously has been suggested to underlie the learning of word forms, i.e.

acoustic-phonetic representations of words (Davis and Gaskell, 2009; Lindsay, 2010). Here, it is proposed that CLS theory also provides the neurobiological basis for word meaning acquisition (Gläser, 2011). More precisely, the initial fast mapping process likely corresponds to the rapid encoding of new word-scene associations in MTL structures, particularly the hippocampus. The localized representations employed by the hippocampus result in context-dependent word meanings similar to those observed in children. It is further proposed that the gradual memory transfer from hippocampal to neocortical sites implements the slow mapping process. During slow mapping, a word's meaning gets decontextualized by abstracting common features among the specific situations in which the word occurred. The reactivation of such situations via the hippocampus enable the neocortex to extract these commonalities via interleaved learning. Since memory transfer and consolidation requires days, weeks, or even months, it explains the prolonged time needed to slow map a word's meaning.

To underpin this proposal multiple hypotheses can be tested. Firstly, CLS theory suggests that an initial encoding of a word meaning primary recruits the hippocampus, whereas the retrieval of this association should successively become independent of the hippocampus later on. A seminal fMRI study on word learning in adults provides evidence in favor of this hypothesis (Breitenstein et al., 2005). The authors showed that the subjects learned consistent, but not inconsistent, picture-pseudoword pairings. More importantly, the initial learning coincided with an increased activity in the MTL structures, particularly the left hippocampus. Subjects who showed larger hippocampal engagement were able to learn the novel vocabulary more efficiently. It could further be shown that hippocampal activity in memory tasks decreased linearly with increasing word proficiency. These results later could be confirmed in experiments that required the subjects to identify word meanings from sentence contexts (Mestres-Missé et al., 2008, 2010).

In addition, CLS theory assigns a pivotal role to sleep, insofar as the generalization and consolidation of word meanings should rely on a sleep-dependent reactivation of memories. Clay et al. (2007) used a variant of the Stroop task to assess the consolidation of word meanings in adults. He could show that picture naming is delayed if a semantically related word is simultaneously presented as compared to a non-related word. This interference effect, however, could only be observed one week after the initial training and hence points to a period of consolidation during word learning. The influence of sleep on infant learning just recently became a research focus (Tarullo et al., 2011). Wilhelm et al. (2008) showed that sleep improves declarative memory performance (e.g. the retrieval of word-pair associations) in 6 to 8-year-old children. More importantly, in another study it could be shown that 15-month-old infants remembered abstract relations between word-pairs, if they slept during a 4 hour interval after training. In contrast, infants that did not sleep only remembered specific word-pair associations (Gómez et al., 2006). This demonstrates that sleep not only facilitates memory consolidation but also memory abstraction. Evidence in favor of the proposal that memory consolidation relies on the reactivation of stored associations comes from two ERP imaging studies (Friedrich and Friederici, 2008, ress). The authors showed that 6- and 14-month-old infants demonstrated the ability to fast map word-object pairs. Importantly, the 6-month-olds showed impaired retention when tested one day after training, whereas the 14-month-olds did not. The reason for this dissociation most probably are the different paces at which MTL structures develop. Whereas a rapid encoding via the hippocampus already might be functional at

birth, it is known that the reactivation of memory items only becomes fully functional approximately at the age of one year. For example, it has been suggested that memory reactivation crucially depends on the dentate gyrus that is immature before this time (Richmond and Nelson, 2007).

In summary, there is ample evidence pointing towards a CLS account for word meaning acquisition. Further hypotheses can be formulated according to this theory. For example, a hippocampal lesion should impair the rapid acquisition of new word meanings, whereas it should not affect retention of previously learned word knowledge. This and other hypotheses remain to be validated by future experiments.

4.1.3. Inferred Computational Principles

Rapid learning and statistical learning do not pose a problem to artificial neural networks (ANNs) if they are considered in isolation. But due to their contradictory computational requirements, a simultaneous application of them constitutes a learning dilemma. According to CLS theory, the problem should not be tackled at the level of a single network but rather at the level of a system (McClelland et al., 1995; O'Reilly and Rudy, 2000; O'Reilly and Norman, 2002). Different components, that are specifically tailored to the contradictory needs of rapid and statistical learning, can serve the purposes of the different tasks. A system-level integration of the components can ultimately produce the desired efficient word learning capability. The CLS account finally boils down to the following three computational principles:

- **Short-term memory (STM) for rapid one-shot learning:**
  New word-scene associations have to be rapidly memorized, i.e. a network has to apply large learning rates. To circumvent catastrophic forgetting the network should comprise localized representations which minimize the interference between the different memory items. This comes at the cost of a decreased generalization ability which qualifies the network as an initial, but temporal, storage site.
- **Long-term memory (LTM) for slow statistical learning:**
  The efficiency of statistically learned representations is what justifies their use for long-term storage. A network that employs statistical learning for the purpose of generalization thereby presupposes overlapping memory representations. This is due to the fact that the representations of different memories require a common basis in order to unveil their commonalities. Since overlapping representations can result in an interference between memory items, small learning rates are essential to prevent catastrophic forgetting.
- **Memory consolidation through tight coupling of STM and LTM:**
  For the purpose of rapid learning novel observations first have to enter the STM network. Here, it is important to note that it is not sufficient for the STM to simply act as a buffer containing the most recent observations. Similar to the LTM, it rather constitutes a fully functional network that allows acquired word knowledge to be used from the very beginning of training. The primary aim of a memory transfer from STM to LTM consequently is to construct a LTM network that is able to memorize the same data as the STM network, but thereby uses more efficient

and more robust representations. A convenient way to realize such a transfer is to reactivate memorized word-scene associations from the STM network and to extract commonalities among them in the LTM network. This means that the STM network internally produces training exemplars and therewith drives the learning in the LTM network.

4.2. Our Computational Model

From a computational point of view, learning a word's meaning corresponds to building a category that encompasses the different referents of the word. A word's category hence separates the situations for which the word provides an appropriate description (the members of the category) from those for which the word does not provide a description (the non-members of the category). Answering the question whether a word can serve as a label for a scene consequently constitutes a binary classification task, insofar as the membership of the scene with respect to the word's category has to be determined. Here, the learning of such categories is considered to be a supervised process, insofar as the words used by a tutor serve as a teaching signal for scene categorization.

As already discussed in Section 2.4, concept formation not only comprises category learning. To efficiently represent a word meaning, a learner additionally has to determine word-relevant feature dimensions on which the category can be built. The relevance of a feature dimension thereby arises from its suitability to discriminate between the members and the non-members of the word category. In the following, a computational model for word meaning acquisition is presented (Gläser and Joublin, 2010b,c). The ability to simultaneously build word meaning categories and extract word-relevant feature dimensions during online operation is what distinguishes the model from previous approaches.

4.2.1. System Architecture

The computational model is largely inspired by CLS theory. It is composed of two complementary but tightly coupled components, which are specifically tailored to achieve a rapid memorization of word-referent pairs and a statistical extraction of commonalities among them. Nevertheless, the model is not meant to provide a 1:1 mapping to certain brain areas; it rather resembles CLS theory from a functional perspective. For this reason, functional correspondences between the model and different brain areas are highlighted in the following. In addition, the description of the framework considers the learning of one category. The meaning of multiple words can be acquired straight-forwardly by using multiple instances of the system.

According to Fig. 4.3 (a), a feature extraction layer first transforms an observation $\boldsymbol{x}$ into a feature pattern $\boldsymbol{y}$. A categorization layer subsequently classifies the feature pattern to either belonging to the category underlying the word's meaning or not. In other words, the membership decision $c \in \{-1, +1\}$ signals whether the word is an appropriate label for the observation $\boldsymbol{x}$. The learning mechanism employed by the framework is illustrated

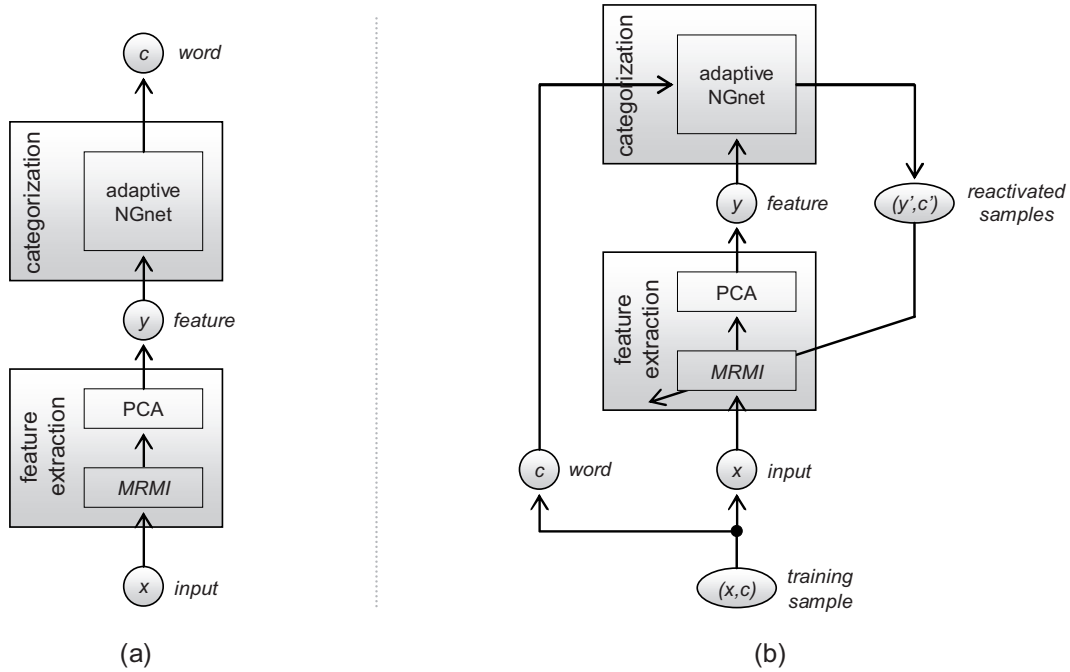


Figure 4.3.: The architecture of the computational model: (a) Input samples $\boldsymbol{x}$ are transformed into feature patterns $\boldsymbol{y}$ which are subsequently categorized. (b) For the learning of word meanings the system components are recurrently coupled.

in Fig. 4.3 (b). It is based on scene-word pairs, formally described by tuples $(\boldsymbol{x}, c)$. For each training sample, learning comprises the execution of the following steps:

1. The categorization layer updates its internal representation according to the training sample. If the sample is not adequately explained by the model, the new association between $\boldsymbol{y}$ and c is memorized. The categorization layer thereby achieves a rapid knowledge acquisition similarly to the hippocampus and further builds a category representation alike multimodal association cortices.

2. The categorization layer generates samples $(\boldsymbol{y}', c')$ according to the internally memorized associations. This process mimics the hippocampal reactivation of patterns during sleep.

3. The generated samples are used to train the feature extraction layer. Statistical learning is applied to extract those feature dimensions that best discriminate the samples $\boldsymbol{y}'$ according to their memberships c'. This type of learning gradually incorporates knowledge on the word category into the extracted features and is related to the interleaved learning process carried out in neocortical areas.

4. The internal representation of the category is adapted to the changed feature dimensions. Due to the fact that the feature space' discriminability is enhanced, it facilitates abstraction in the categorization layer. More precisely, the different referents of a word should be more closely located in the new feature space, whereas they should be largely separated from the non-members of the category. This finally allows the categorization layer to construct a more efficient and more robust representation of the word category.

These steps can be iteratively applied whenever a new training sample arrives. By doing so, the knowledge about the learned category gradually shifts from the categorization layer into the extracted features. As a consequence, the features more and more facilitate the classification task as training progresses and hence drive the generalization in the categorization layer. This gradual memory transfer and generalization is akin to the consolidation dynamics suggested by CLS theory (cf. Fig. 4.2). However, the framework does not draw an explicit distinction between short- and long-term memory as CLS theory does; it rather uses the same network (the categorization layer) for both purposes. This is possible, since the applied network possesses dedicated learning mechanisms that satisfy the needs of *rapid learning using localized representations* and *statistical learning using overlapping representations* at the same time. In the following, detailed descriptions of the system components are provided.

4.2.2. Categorization Layer

The aim of the classification layer is to provide a classification function $\widetilde{c} = \Omega(\boldsymbol{y})$ such that Ω is an approximation of the mapping $\mathcal{S}_\mathcal{Y} \mapsto [-1, +1]$, where $\mathcal{S}_\mathcal{Y}$ denotes the feature space encompassing all observations and $[-1, +1]$ the word category membership. As Ω has to be invertible, i.e. $\Omega^{-1} : [-1, +1] \mapsto \mathcal{S}_\mathcal{Y}$, to enable a reactivation of memorized associations between feature patterns $\boldsymbol{y}$ and category memberships c, modelling the classification function with a generative model is best suited to our task. Here, a *Normalized Gaussian Network (NGnet)* is used for this purpose.

Normalized Gaussian Network

An NGnet (Moody and Darken, 1989) can serve as a universal function approximator. For the sake of generality, in the following we consider the approximation of mappings $\Omega : \mathbb{R}^{\mathcal{D}_y} \mapsto \mathbb{R}^{\mathcal{D}_c}$ from an $\mathcal{D}_y$-dimensional input space to a $\mathcal{D}_c$-dimensional output space (even though $\mathcal{D}_c = 1$ in our application, since memberships to only one word category are considered). Given an input $\boldsymbol{y}$ an NGnet's output $\widetilde{\boldsymbol{c}}(\boldsymbol{y})$ is calculated according to

$$\widetilde{\boldsymbol{c}}(\boldsymbol{y}) = \frac{1}{\sum_{j=1}^{M} \phi_j(\boldsymbol{y})} \cdot \sum_{i=1}^{M} \boldsymbol{\alpha}_i \cdot \phi_i(\boldsymbol{y}). \tag{4.1}$$

Thereby, $\phi_i(\boldsymbol{y})$ denotes the response of the i-th hidden unit to input $\boldsymbol{y}$, M is the number of hidden units, and $\boldsymbol{\alpha}_i$ the weight vector from unit i to the output neurons (see Fig. 4.4). For the purpose of calculating category memberships, the continuously valued output $\widetilde{\boldsymbol{c}}(\boldsymbol{y})$ can be converted to a binary decision ($[-1, +1]$) via thresholding with the sign function, i.e. $\widehat{\boldsymbol{c}}(\boldsymbol{y}) = \text{sign}(\widetilde{\boldsymbol{c}}(\boldsymbol{y}))$.

The NGnet uses localized representations, insofar as the response of a hidden unit i is described by a multivariate Gaussian of form

$$\phi_i(\boldsymbol{y}) = \exp\left(-\frac{1}{2} \cdot (\boldsymbol{y} - \boldsymbol{\mu_i})^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu_i})\right), \tag{4.2}$$

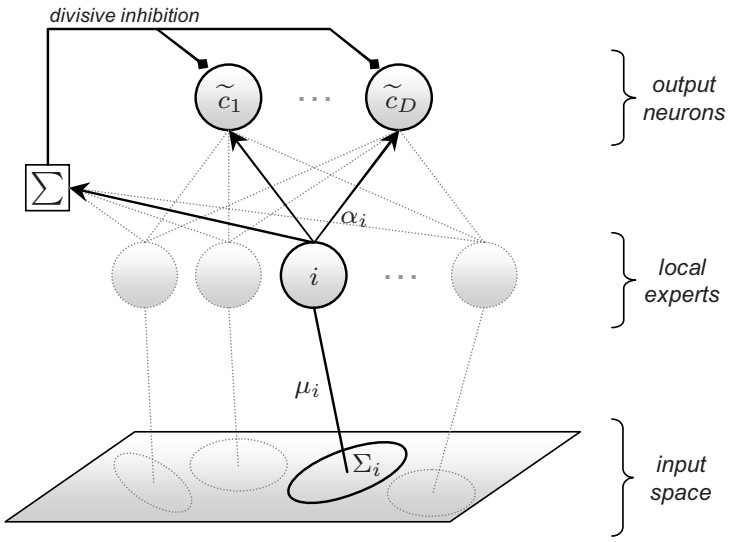


Figure 4.4.: The architecture of an NGnet.

where $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma}_i$ denote the center and covariance matrix of the Gaussian. Therefore, an NGnet is similar to a standard RBF network except for the normalization of the output by the total hidden unit activity (cf. Eq. (4.1)). The effect of this normalization is a competition between the hidden units. More precisely, it is a competition for responsibility in representing the inputs $\boldsymbol{y}$. The hidden units softly partition the input space $\mathcal{S}_{\mathcal{Y}}$, such that each unit is responsible for inputs stemming from its associated input region. As exemplarily depicted in Fig. 4.5, the hidden units consequently constitute local models (or local experts (Jacobs et al., 1991)) of the mapping to be approximated. Each local expert provides a constant approximation of the target function which is valid for the restricted input region the expert is responsible for. The overall approximation is finally obtained by overlaying (or weighting) the local approximations. A further advantage of the normalization term is an improved inter- and extrapolation compared to RBF networks. This is also illustrated in Fig. 4.5. The weighting extends local approximations also to such regions, that are not covered by hidden units. An NGnet hence constitutes an exemplar-based network suitable for similarity-based generalization.

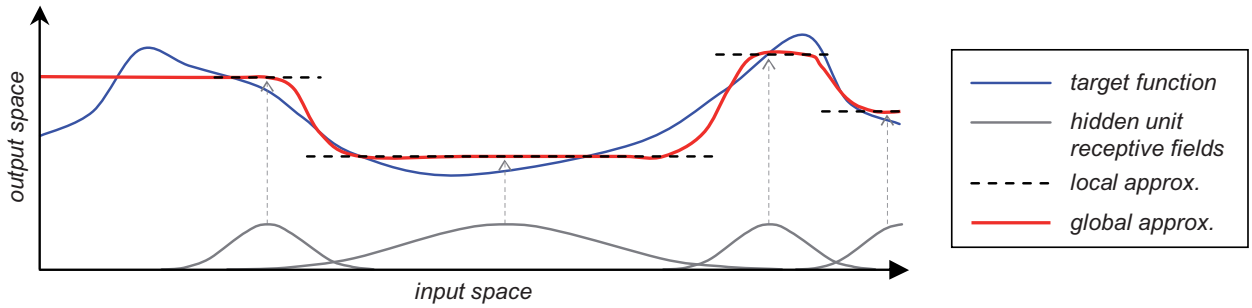


Figure 4.5.: The application of an NGnet to a one-dimensional function approximation problem: The hidden units locally approximate the target function by constant values. A global approximation is obtained by weighting the local approximations according to the input regions the hidden units are responsible for.

Probabilistic Interpretation & Online Training

Xu (1998) presented a stochastic interpretation of NGnets. He showed how this probabilistic view can be used to train the parameters of the network during online operation. The approach of Xu is one of the learning mechanisms applied in our computational model. It further provides the basis for additional learning processes that will be presented later in this section. For this reason, the main aspects of the work of Xu will be reviewed in the following.

Consider an NGnet to be a generative model. A sample $(\boldsymbol{y}, \boldsymbol{c})$ thereby is a stochastic event that has been generated from one of the network's local experts. Therefore, let each local expert be fully described by two Gaussian probability density functions (pdfs), one over the input space and one over output space:

$$p(\boldsymbol{y}|i, \boldsymbol{\Theta}) = G(\boldsymbol{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4.3}$$

$$p(\boldsymbol{c}|i, \boldsymbol{\Theta}) = G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i) \tag{4.4}$$

Here, $\boldsymbol{\Theta} = \{\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i\}_{i=1}^{M}\}$ denotes the parameters of the NGnet and $G(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{S})$ the multivariate normal distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{S}$ evaluated at $\boldsymbol{x} \in \mathbb{R}^K$:

$$G(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{S}) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{S}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right). \tag{4.5}$$

Both pdfs are illustrated in Fig. 4.6. The Gaussian over the input space corresponds to the receptive field of an expert i, where μ_i and Σ_i refer to its location and size, respectively. Hence, $p(\boldsymbol{y}|i, \boldsymbol{\Theta})$ denotes the probability that an expert i generates an input $\boldsymbol{y}$. The Gaussian over the output space is described by its mean α_i, which corresponds to the constant value that is used to locally approximate the target function. Furthermore, it is described by its covariance matrix Γ_i, which provides a measure for the quality of this approximation. This means that a wide Gaussian refers to a bad (or uncertain) approximation and vice versa. The value $p(\boldsymbol{c}|i, \boldsymbol{\Theta})$ consequently measures the probability that an output $\boldsymbol{c}$ is drawn from this pdf.

Due to the fact that the receptive fields of the local experts partly overlap, an input $\boldsymbol{y}$ can be generated from multiple experts. The experts consequently compete for responsibility

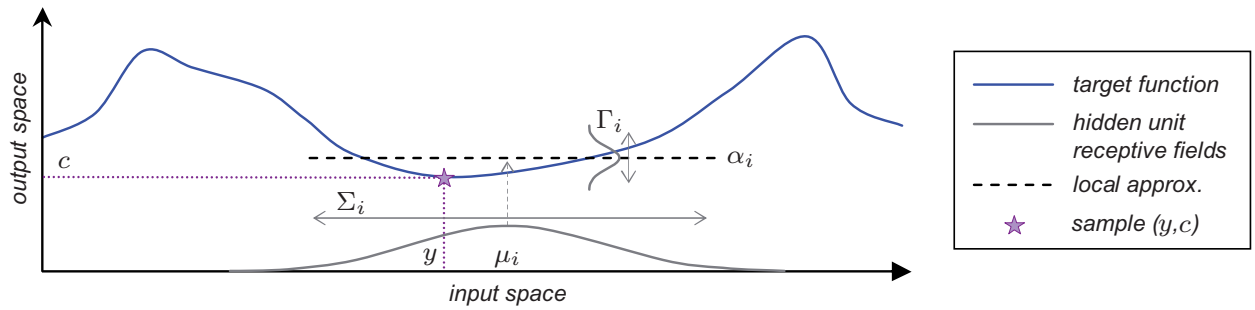


Figure 4.6.: An illustration of the Gaussian pdfs used for the description of a local model i (see text for details).

in representing an input. According to Xu (1998) this competition can be modeled via weights $a_i = |\mathbf{\Sigma}_i|^{1/2} / \sum_{j=1}^{M} |\mathbf{\Sigma}_j|^{1/2}$. Then the probability that the hidden unit i is responsible for the input $\boldsymbol{y}$ turns out to be the ratio of the i-th unit's response to the input and the total hidden layer activity:

$$\begin{aligned} p(i|\boldsymbol{y}, \mathbf{\Theta}) &= \frac{a_i \cdot p(\boldsymbol{y}|i, \mathbf{\Theta})}{\sum_{j=1}^{M} a_j \cdot p(\boldsymbol{y}|j, \mathbf{\Theta})} \\ &= \frac{\frac{|\mathbf{\Sigma}_i|^{1/2}}{\sum_{j=1}^{M} |\mathbf{\Sigma}_j|^{1/2}} \cdot \frac{1}{(2\pi)^{\mathcal{D}_y/2} |\mathbf{\Sigma}_i|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\boldsymbol{y} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_i)\right)}{\sum_{j=1}^{M} \frac{|\mathbf{\Sigma}_j|^{1/2}}{\sum_{m=1}^{M} |\mathbf{\Sigma}_m|^{1/2}} \cdot \frac{1}{(2\pi)^{\mathcal{D}_y/2} |\mathbf{\Sigma}_j|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\boldsymbol{y} - \boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_j)\right)} \\ &= \frac{\exp\left(-\frac{1}{2} \cdot (\boldsymbol{y} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_i)\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{2} \cdot (\boldsymbol{y} - \boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_j)\right)} \\ &= \frac{\phi_i(\boldsymbol{y})}{\sum_{j=1}^{M} \phi_j(\boldsymbol{y})}. \end{aligned} \tag{4.6}$$

The posterior probability $p(\boldsymbol{c}|\boldsymbol{y}, \mathbf{\Theta})$ that an output $\boldsymbol{c}$ is generated given an input $\boldsymbol{y}$ consequently can be calculated according to

$$\begin{aligned} p(\boldsymbol{c}|\boldsymbol{y}, \mathbf{\Theta}) &= \sum_{i=1}^{M} p(i|\boldsymbol{y}, \mathbf{\Theta}) \cdot p(\boldsymbol{c}|i, \mathbf{\Theta}) \\ &= \sum_{i=1}^{M} \frac{\phi_i(\boldsymbol{y})}{\sum_{j=1}^{M} \phi_j(\boldsymbol{y})} \cdot G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \mathbf{\Gamma}_i). \end{aligned} \tag{4.7}$$

Then, the expected output of the generative model is

$$E(\boldsymbol{c}|\boldsymbol{y}, \mathbf{\Theta}) = \frac{1}{\sum_{j=1}^{M} \phi_j(\boldsymbol{y})} \cdot \sum_{i=1}^{M} \boldsymbol{\alpha}_i \cdot \phi_i(\boldsymbol{y}). \tag{4.8}$$

As can be seen, the conditional expectation $E(\boldsymbol{c}|\boldsymbol{y}, \mathbf{\Theta})$ matches the definition of an NGnet (cf. (Eq. 4.1)). Hence, the parameters $\mathbf{\Theta}$ of the NGnet can be estimated by maximum likelihood learning on the log-likelihood of the observed data $(\{\boldsymbol{y}\}, \{\boldsymbol{c}\})$. Thereby, online learning necessitates an algorithm that adapts the network parameters $\mathbf{\Theta}$ sequentially whenever a new training sample is obtained. This can be achieved by an iterative application of Expectation-Maximization (EM). Using stochastic approximation of the form

$$\mathbf{\Theta}(t) = \mathbf{\Theta}(t-1) + \eta \cdot \frac{\partial e_t}{\partial \mathbf{\Theta}(t-1)}, \tag{4.9}$$

where η is a learning rate and e_t the error for the t-th sample, the sequential EM algorithm proposed by Xu (1998) can be summarized as follows:

- **E-step:** Given the current estimator value $\boldsymbol{\Theta}(t-1)$, the posterior probability $p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}(t-1))$ of assigning the t-th training sample $(\boldsymbol{y}_t, \boldsymbol{c}_t)$ to the i-th local model can be calculated according to Bayes rule:

$$\begin{aligned} p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}(t-1)) &= \frac{p(i|\boldsymbol{y}_t, \boldsymbol{\Theta}(t-1)) \cdot p(\boldsymbol{c}_t|i, \boldsymbol{\Theta}(t-1))}{p(\boldsymbol{c}_t|\boldsymbol{y}_t, \boldsymbol{\Theta}(t-1))} \\ &= \frac{\phi_i(\boldsymbol{y}_t) \cdot G(\boldsymbol{c}_t, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i)}{\sum_{j=1}^{M} \phi_j(\boldsymbol{y}_t) \cdot G(\boldsymbol{c}_t, \boldsymbol{\alpha}_j, \boldsymbol{\Gamma}_j)} \end{aligned} \tag{4.10}$$

- **M-step:** The log-likelihood becomes maximized by calculating $\boldsymbol{\Theta}(t)$ with

$$\begin{aligned} \boldsymbol{\mu}_i(t) &= \boldsymbol{\mu}_i(t-1) + \eta_i(t) \cdot (\boldsymbol{y}_t - \boldsymbol{\mu}_i(t-1)) \\ \boldsymbol{\Sigma}_i(t) &= (1-\eta_i(t)) \cdot \boldsymbol{\Sigma}_i(t-1) + \eta_i(t) \cdot [\boldsymbol{y}_t - \boldsymbol{\mu}_i(t-1)][\boldsymbol{y}_t - \boldsymbol{\mu}_i(t-1)]^T \\ \boldsymbol{\alpha}_i(t) &= \boldsymbol{\alpha}_i(t-1) + \eta_i(t) \cdot (\boldsymbol{c}_t - \boldsymbol{\alpha}_i(t-1)) \\ \boldsymbol{\Gamma}_i(t) &= (1-\eta_i(t)) \cdot \boldsymbol{\Gamma}_i(t-1) + \eta_i(t) \cdot [\boldsymbol{c}_t - \boldsymbol{\alpha}_i(t-1)][\boldsymbol{c}_t - \boldsymbol{\alpha}_i(t-1)]^T. \end{aligned} \tag{4.11}$$

  Thereby, the posteriors $p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}(t-1))$ enter the M-Step via adaptive learning rates η_i that are individually calculated for each local model i:

$$\begin{aligned} \eta_i(t) &= \eta \cdot \frac{p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}(t-1))}{\gamma_i(t-1)} \\ n_i(t) &= (1-\eta_i(t)) \cdot n_i(t-1) + \eta_i(t) \\ \gamma_i(t) &= \frac{n_i(t)}{\sum_{j=1}^{M} n_j(t)}. \end{aligned} \tag{4.12}$$

  Here, η is a baseline learning rate.

The calculation of the individual learning rates is interesting as it implements a kind of homeostasis. As can be seen from Eq. (4.12), $n_i(t)$ calculates a running average of the instantaneous learning rates $\eta_i(t)$. By normalizing this average learning rate, $\gamma_i(t)$ represents the i-th model's proportion on the overall learning in the network. That means that $\gamma_i(t)$ will decrease if no training samples are assigned to the i-th model, since this would results in near-zero instantaneous learning rates (due to very small posteriors $p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}(t-1))$). By scaling the instantaneous learning rates $\eta_i(t)$ with the inverse of $\gamma_i(t)$ this effect is circumvented. As a result, a local model, that is responsible for many training samples, will incorporate these samples with relatively small instantaneous learning rates. In contrast, a local model, that is responsible for just a few training samples, will learn on these samples with relatively high instantaneous learning rates. As training progresses, the local models thus will approach an approximately similar average plasticity.

The reviewed algorithm shows how network parameters can be trained during online operation. However, two further issues remain to be solved. Firstly, training a network via EM is a statistical learning approach that requires much training data. Additional mechanisms are consequently needed to achieve a rapid learning from few training exemplars. Secondly, it is unclear how many hidden units the network should contain. In the following, a solution to both aspects is presented, insofar as the NGnet is extended by local model manipulation mechanisms (Gläser and Joublin, 2010a).

Local Model Manipulation Mechanisms

One of the main problems when using an NGnet is the specification of the network's complexity, i.e. the selection of the number of hidden units (local experts). Solving this problem is usually done by incorporating domain knowledge. Difficult approximation problems will obviously necessitate more hidden units than simple tasks. However, it is desirable to build general purpose network models which are able to autonomously adapt their complexity based on the problem at hand. For an NGnet, this involves mechanisms for assigning new local experts and removing, splitting, or merging existing ones. Furthermore, criteria for deciding when to execute the model manipulation mechanisms have to be defined.

In previous work several methods for an incremental build-up of an NGnet have been proposed (Samejima and Omori, 1999; Lu et al., 1997; Huang et al., 2005; Sekino et al., 2005; Sato and Ishii, 2000). Some of them implement algorithms which are in part similar to the ones that are outlined next. However, these methods usually assume the complexity of the task to be constant over time. Consequently, they increase an NGnet's complexity until a sufficient approximation quality is achieved. In contrast, we explicitly take into account a varying task complexity and, thus, present mechanisms which continuously adapt an NGnet's complexity according to task demands. As one example, we introduce the merging of experts which turns out to be beneficial for obtaining small-sized networks and improving generalization (see results in Section 4.3).

Model Removal

Local experts with little or no contribution to an NGnet's approximation are redundant and should be removed. The posterior $p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta})$ refers to the probability of assigning the t-th sample $(\boldsymbol{y}_t, \boldsymbol{c}_t)$ to the i-th local expert (see Eq. 4.10). Let ρ_i denote the running average over this posterior, where η is a time constant:

$$\rho_i(t) = (1 - \eta) \cdot \rho_i(t-1) + \eta \cdot p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta}). \tag{4.13}$$

Since the posteriors $p(i|\boldsymbol{y}_t, \boldsymbol{c}_t, \boldsymbol{\Theta})$ are typically either close to 1 or close to 0, ρ_i is proportional to the average number of samples for which the local model i best describes the mapping task. Consequently, ρ_i measures the contribution of the i-th expert to the overall network output. It hence serves as an importance weight such that $\rho_i < \theta_{remove}/M$ with $0 < \theta_{remove} \ll 1$ constitutes a criterion for removing the i-th local expert. Thereby, M denotes the number of models.

Let $\mathcal{M}_i$ denote the i-th expert and $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_i\}_{i=1}^{M}$ the set of all local experts. When removing the i-th model, we also adapt the ρ_j such that $\sum_j \rho_j$ before and after the removal remains unchanged, i.e.

$$\begin{aligned} \boldsymbol{\mathcal{M}}^* &= \boldsymbol{\mathcal{M}} \backslash \{\mathcal{M}_i\} \\ \rho_j^* &= \frac{|\boldsymbol{\mathcal{M}}|}{|\boldsymbol{\mathcal{M}}^*|} \cdot \rho_j \qquad ,\forall j \text{ with } \mathcal{M}_j \in \boldsymbol{\mathcal{M}}^*. \end{aligned} \tag{4.14}$$

Here, $|\boldsymbol{\mathcal{S}}|$ denotes the cardinality of the set $\boldsymbol{\mathcal{S}}$.

Model Assignment

A new local expert should be assigned, if a training sample $(\boldsymbol{y}_t, \boldsymbol{c}_t)$ is novel or surprising to the network. Thereby, novelty refers to the fact that the sample (or a similar one) has not been observed before and hence is not sufficiently well covered by any of the existing local experts. In contrast, surprise relates to a large deviation of the supplied output $\boldsymbol{c}_t$ from the expected one (the network's approximation $\widetilde{\boldsymbol{c}}(\boldsymbol{y}_t)$). These criteria can be expressed as follows

$$\max_i p(\boldsymbol{y}_t, \boldsymbol{c}_t | i, \boldsymbol{\Theta}) < \theta_{coverage} \tag{4.15}$$

$$e_t > \theta_{surprise}, \tag{4.16}$$

where $\theta_{coverage}$ and $\theta_{surprise}$ are thresholds and

$$\begin{aligned} p(\boldsymbol{y}_t, \boldsymbol{c}_t | i, \boldsymbol{\Theta}) &= p(\boldsymbol{c}_t | i, \boldsymbol{\Theta}) \cdot p(\boldsymbol{y}_t | i, \boldsymbol{\Theta}) \\ &= G(\boldsymbol{c}_t, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i) \cdot G(\boldsymbol{y}_t, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad (4.17) \\ e_t &= [\boldsymbol{c}_t - \widetilde{\boldsymbol{c}}(\boldsymbol{y}_t)]^T [\boldsymbol{c}_t - \widetilde{\boldsymbol{c}}(\boldsymbol{y}_t)]. \qquad (4.18) \end{aligned}$$

As can be seen, the coverage is measured according to how good the sample fits the experts' Gaussian pdfs over the input and output space. In contrast, the squared approximation error serves as a measure for the surprise. If any of these conditions is fulfilled, a new local model is added to the NGnet and the importance weights ρ_j are adapted, such that $\sum_j \rho_j$ before and after the assignment remain unchanged:

$$\begin{aligned} \boldsymbol{\mathcal{M}}^* &= \boldsymbol{\mathcal{M}} \cup \{\mathcal{M}_{new}\} \\ \rho_j^* &= \frac{|\boldsymbol{\mathcal{M}}|}{|\boldsymbol{\mathcal{M}}^*|} \cdot \rho_j \qquad , \forall j \text{ with } \mathcal{M}_j \in \boldsymbol{\mathcal{M}}. \end{aligned} \tag{4.19}$$

The new model $\mathcal{M}_{new}$ can be initialized as proposed by Sato and Ishii (2000). As illustrated in Fig. 4.7 (a), its receptive field is centered at the input $\boldsymbol{y}_t$, whereas the size of the receptive field is determined by the distance to the closest existing model. The new model produces a local approximation $\boldsymbol{\alpha}_{new}$ of the target function that equals the observed output $\boldsymbol{c}_t$. Thereby, the uncertainty about this approximation, i.e. the covariance of the Gaussian pdf over the output space, is initialized to the maximum uncertainty of the existing models.

$$\begin{aligned} \boldsymbol{\mu}_{new} &= \boldsymbol{y}_t \\ \boldsymbol{\Sigma}_{new} &= \min_i \left(\frac{[\boldsymbol{\mu}_i - \boldsymbol{\mu}_{new}]^T [\boldsymbol{\mu}_i - \boldsymbol{\mu}_{new}]}{\mathcal{D}_y} \right) \cdot \boldsymbol{I} \\ \boldsymbol{\alpha}_{new} &= \boldsymbol{c}_t \\ \boldsymbol{\Gamma}_{new} &= \max_i \boldsymbol{\Gamma}_i \\ \rho_{new} &= \frac{1}{|\boldsymbol{\mathcal{M}}| + 1} \end{aligned} \tag{4.20}$$

Here, $\boldsymbol{I}$ denotes the identity matrix and $\mathcal{D}_y$ the input dimensionality. In the special case of assigning the first local model, $\boldsymbol{\Sigma}_{new}$ and $\boldsymbol{\Gamma}_{new}$ are initialized to some predefined $\boldsymbol{\Sigma}_{init}$ and $\boldsymbol{\Gamma}_{init}$, respectively, as well as $\rho_{new} = 1$.

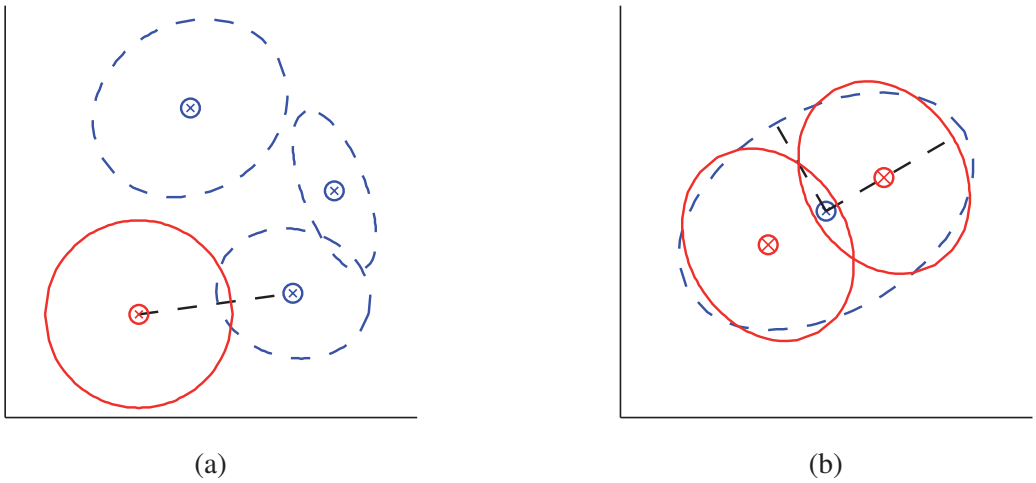


Figure 4.7.: An illustration of model assignment and model splitting: In (a) a new model is allocated and initialized according to its minimum distance to already existing models. In (b) a model is split along its principal dimension.

Model Splitting

If the i-th local model's quality of approximating the mapping task is insufficient, the input space region corresponding to its receptive field should be refined and covered by multiple experts. An insufficient approximation quality is characterized by a diffuse probability distribution $p(\boldsymbol{c}|i, \boldsymbol{\Theta}) = G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i)$ over the output space. Consequently, the size of the Gaussian (for which $|\boldsymbol{\Gamma}_i|$ is an indicator) is an appropriate criterion for splitting a model. In summary, the i-th model is split if

$$|\boldsymbol{\Gamma}_i| > \theta_{split}, \tag{4.21}$$

where θ_{split} is a threshold. If this criterion is met, $\mathcal{M}_i$ is adjusted to $\mathcal{M}_i^*$, a new model $\mathcal{M}_{new}$ is created, and finally added to the model pool.

$$\boldsymbol{\mathcal{M}}^* = (\boldsymbol{\mathcal{M}} \backslash \mathcal{M}_i) \cup \{\mathcal{M}_i^*, \mathcal{M}_{new}\} \tag{4.22}$$

The splitting is performed along the prominent dimension of the receptive field, which is similar to the method proposed by Samejima and Omori (1999). Therefore, let $\boldsymbol{\zeta}_n$ and κ_n denote the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}_i$ sorted in descending order of the eigenvalues, i.e. $\kappa_1 \geq \kappa_2 \geq \ldots \geq \kappa_{\mathcal{D}_y}$. The spin-off models $\mathcal{M}_i^*$ and $\mathcal{M}_{new}$ are initialized as

$$\boldsymbol{\mu}_i^*, \boldsymbol{\mu}_{new} = \boldsymbol{\mu}_i \pm \xi_1 \cdot \sqrt{\kappa_1} \cdot \boldsymbol{\zeta}_1 \tag{4.23}$$

$$\boldsymbol{\alpha}_i^*, \boldsymbol{\alpha}_{new} = \boldsymbol{\alpha}_i \tag{4.24}$$

$$\boldsymbol{\Sigma}_i^*, \boldsymbol{\Sigma}_{new} = \frac{\xi_2}{\kappa_1} \cdot \boldsymbol{\zeta}_1 \boldsymbol{\zeta}_1^T + \sum_{n=2}^{\mathcal{D}_y} \frac{1}{\kappa_n} \cdot \boldsymbol{\zeta}_n \boldsymbol{\zeta}_n^T \tag{4.25}$$

$$\boldsymbol{\Gamma}_i^*, \boldsymbol{\Gamma}_{new} = 0.5 \cdot \boldsymbol{\Gamma}_i \tag{4.26}$$

$$\rho_i^*, \rho_{new} = 0.5 \cdot \rho_i, \tag{4.27}$$

where ξ_1 and ξ_2 are constants controlling the overlap of their receptive fields. The splitting mechanism is illustrated in Fig. 4.7 (b).

Model Merging

If multiple local models are sufficiently similar, they can be merged to one local expert. The similarity depends on the overlap between the experts' pdfs over the input and output space, respectively. Let $\mathcal{U}(\mathcal{M}_i, \mathcal{M}_j)$ be a function measuring the similarity between two local models $\mathcal{M}_i$ and $\mathcal{M}_j$ with $0 \leq \mathcal{U}(\mathcal{M}_i, \mathcal{M}_j) \leq 1$. Thereby, a value of 1 corresponds to model identity and a value of 0 to total model dissimilarity. Furthermore, let $\mathcal{V}(p(\boldsymbol{a}), q(\boldsymbol{a}))$ be a function measuring the overlap between two multivariate pdfs $p(\boldsymbol{a})$ and $q(\boldsymbol{a})$ with $0 \leq \mathcal{V}(p(\boldsymbol{a}), q(\boldsymbol{a})) \leq 1$. Then we define

$$\begin{aligned} \mathcal{U}(\mathcal{M}_i, \mathcal{M}_j) &= \mathcal{V}(\, p(\boldsymbol{y}|i, \boldsymbol{\Theta})\,,\, p(\boldsymbol{y}|j, \boldsymbol{\Theta})\,) \cdot \mathcal{V}(\, p(\boldsymbol{c}|i, \boldsymbol{\Theta})\,,\, p(\boldsymbol{c}|j, \boldsymbol{\Theta})\,) \\ &= \mathcal{V}(\, G(\boldsymbol{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\,,\, G(\boldsymbol{y}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\,) \cdot \mathcal{V}(\, G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i)\,,\, G(\boldsymbol{c}, \boldsymbol{\alpha}_j, \boldsymbol{\Gamma}_j)\,). \end{aligned} \tag{4.28}$$

Consequently, measuring the pair-wise similarity between local experts reduces to calculating the overlap between multivariate Gaussian pdfs. The *Bhattacharyya Coefficient (BC)* provides an approximation for this. It is defined as

$$BC(p, q) = \int_{\boldsymbol{a}} \sqrt{p(\boldsymbol{a}) \cdot q(\boldsymbol{a})}\, d\boldsymbol{a}, \tag{4.29}$$

for which a closed form solution exists for multivariate Gaussians $p(\boldsymbol{a}) = G(\boldsymbol{a}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $q(\boldsymbol{a}) = G(\boldsymbol{a}, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$:

$$D_B(p, q) = \frac{1}{8} \cdot (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T * \boldsymbol{\Sigma}^{-1} * (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q) + \frac{1}{2} \cdot \log \frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma}_p| \cdot |\boldsymbol{\Sigma}_q|}} \tag{4.30}$$

$$BC(p, q) = \exp(-D_B(p, q)). \tag{4.31}$$

Here, D_B is called the *Bhattacharyya distance* with $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q)/2$. In summary, the similarity $\mathcal{U}(\mathcal{M}_i, \mathcal{M}_j)$ between two local models is calculated according to Eq. (4.28), where we set $\mathcal{V}(p, q) = BC(p, q)$. If the similarity exceeds a threshold θ_{merge}, i.e.

$$\mathcal{U}(\mathcal{M}_i, \mathcal{M}_j) > \theta_{merge}, \tag{4.32}$$

the models $\mathcal{M}_i$ and $\mathcal{M}_j$ are merged. The NGnet is finally adapted such that

$$\boldsymbol{\mathcal{M}}^* = (\boldsymbol{\mathcal{M}} \backslash \{\mathcal{M}_i, \mathcal{M}_j\}) \cup \{\mathcal{M}_{new}\}. \tag{4.33}$$

The creation of the new model $\mathcal{M}_{new}$ involves two steps. Firstly, the importance weight of the new model ρ_{new} is set to $\rho_{new} = \rho_i + \rho_j$. Secondly, the multivariate Gaussian pdfs of the models $\mathcal{M}_i$ and $\mathcal{M}_j$ are merged. More precisely, the pdfs over the input space ($G(\boldsymbol{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $G(\boldsymbol{y}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$) are merged to $G(\boldsymbol{y}, \boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})$. Similarly, the pdfs over the output space ($G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i)$ and $G(\boldsymbol{c}, \boldsymbol{\alpha}_j, \boldsymbol{\Gamma}_j)$) are merged to $G(\boldsymbol{c}, \boldsymbol{\alpha}_{new}, \boldsymbol{\Gamma}_{new})$. Thereby, those local models, that possess large importance weights ρ, dominate the merging over less important ones. Merging the pdfs over the input space consequently aims at minimizing the functional $\mathcal{F}$ with

$$\mathcal{F} = \sum_{r \in \{i,j\}} \omega_r \cdot D(\, G(\boldsymbol{a}, \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)\, ||\, G(\boldsymbol{a}, \boldsymbol{\mu}_{new}, \boldsymbol{\Sigma}_{new})\,), \tag{4.34}$$

where D is a divergence measure and $\omega_r = \rho_r/(\rho_i + \rho_j)$ are the normalized importance weights with $r \in \{i, j\}$. The creation of $G(\boldsymbol{c}, \boldsymbol{\alpha}_{new}, \boldsymbol{\Gamma}_{new})$ follows a similar minimization problem.

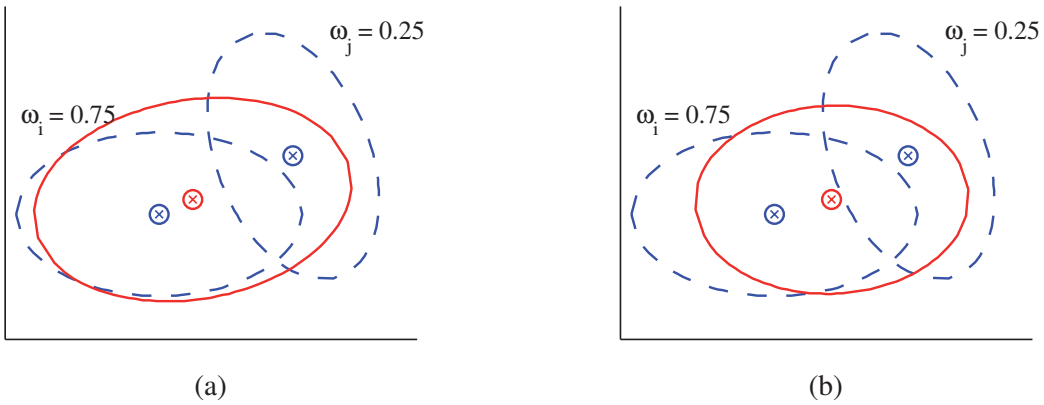


Figure 4.8.: The merging of two 2-dimensional Gaussian distributions (dashed lines) with importance weights $\omega_i = 0.75$ and $\omega_j = 0.25$ is illustrated: (a) shows the result when the Kullback-Leibler divergence is used for clustering, whereas (b) depicts the result when using the Jenson-Shannon divergence.

Existing approaches for minimizing $\mathcal{F}$ mainly differ in the used divergence measure D. Two of them are of particular interest: Firstly, *Kullback-Leibler divergence* based clustering (Davis and Dhillon, 2006) and, secondly, *Jenson-Shannon divergence* based clustering (Myrvoll and Soong, 2003). Both approaches are reviewed in Appendix A, where a detailed derivation of formulas for the calculation of the resulting Gaussians is presented. Here, the results of both methods are just exemplarily depicted in Fig. 4.8. As can be seen, the divergence measures result in different clusters. More precisely, the Gaussian obtained via Kullback-Leiber divergence based clustering is larger than the one obtained by Jenson-Shannon divergence based clustering. The former is nearly the union of the individual Gaussians. Thus, Kullback-Leibler divergence based clustering seems to be the appropriate technique when the receptive fields of Gaussians should be joined. However, it is inappropriate for joining (normalized) probability distributions, for which Jenson-Shannon divergence-based clustering yields better results. Since the competition between the local experts of an NGnet overwrites the normalization of $p(\boldsymbol{y}|i, \boldsymbol{\Theta})$ (cf. Eq. (4.10)), Kullback-Leibler divergence-based clustering is used to construct $G(\boldsymbol{y}, \boldsymbol{\mu_{new}}, \boldsymbol{\Sigma}_{new})$. In contrast, Jenson-Shannon divergence-based clustering is used for calculating $G(\boldsymbol{c}, \boldsymbol{\alpha_{new}}, \boldsymbol{\Gamma}_{new})$. In summary, the merging of local models can be done using the greedy strategy depicted in Algorithm 4.1.

Algorithm 4.1 Merge Local Models

Calculate the similarity $\mathcal{U}(\mathcal{M}_i, \mathcal{M}_j), \quad \forall\{\mathcal{M}_i, \mathcal{M}_j\}$
$\{a, b\} \leftarrow \arg\max_{\{i,j\}} \mathcal{U}(\mathcal{M}_i, \mathcal{M}_j)$
while $\mathcal{U}(\mathcal{M}_a, \mathcal{M}_b) > \theta_{merge}$ **do**
  Merge $\mathcal{M}_a$ and $\mathcal{M}_b$ to $\mathcal{M}_{new}$
  $\boldsymbol{\mathcal{M}} \leftarrow (\boldsymbol{\mathcal{M}} \backslash \{\mathcal{M}_a, \mathcal{M}_b\}) \cup \{\mathcal{M}_{new}\}$
  Update the similarity $\mathcal{U}(\mathcal{M}_i, \mathcal{M}_j), \quad \forall\{\mathcal{M}_i, \mathcal{M}_j\}$
  $\{a, b\} \leftarrow \arg\max_{\{i,j\}} \mathcal{U}(\mathcal{M}_i, \mathcal{M}_j)$
end while

Incremental Learning Algorithm

The local model manipulation mechanisms can be combined with online EM training as shown in Algorithm 4.2. This results in an adaptive version of an NGnet, insofar as the network can grow and shrink according to the demands of a task. It increases its complexity by allocating or splitting hidden units until a sufficient approximation quality is reached. On the contrary, the network strives for a compact size by removing or merging redundant units. The adaptive NGnet further enables rapid learning. This is achieved by assigning a new hidden unit whenever a novel or surprising training sample is encountered. The new unit memorizes the observed pattern in one-shot. Since the adaptive NGnet also applies EM training, the network resembles rapid hippocampal learning as well as a slower statistical learning. The latter might be carried out in multimodal association areas surrounding the hippocampus (e.g. PHC or PRC; cf. Fig 4.1).

Algorithm 4.2 Adaptive NGnet

Initialize $\boldsymbol{\mathcal{M}} \leftarrow \emptyset$

for all samples $(\boldsymbol{y}_t, \boldsymbol{c}_t)$ **do**

if $\boldsymbol{\mathcal{M}} = \emptyset$ **then**

assign a new model

else

{Sample Coverage & Surprise}
$coverage \leftarrow \max_i p(\boldsymbol{y}_t, \boldsymbol{c}_t | i, \boldsymbol{\Theta}(t-1))$
$surprise \leftarrow [\boldsymbol{c}_t - \widetilde{\boldsymbol{c}}(\boldsymbol{y}_t)]^T [\boldsymbol{c}_t - \widetilde{\boldsymbol{c}}(\boldsymbol{y}_t)]$

{Network Training}
if $(coverage < \theta_{coverage})$ **or** $(surprise > \theta_{surprise})$ **then**
assign a new model
else
train the NGnet via sequential EM learning
end if

{Local Model Manipulation}
remove all models $\mathcal{M}_i$ with $\rho_i < \theta_{remove}$
split all models $\mathcal{M}_i$ with $|\boldsymbol{\Gamma}_i| > \theta_{split}$
merge models using Algorithm 4.1

end if

end for

4.2.3. Feature Extraction Layer

The aim of the feature extraction layer is to provide a transformation Φ that maps inputs $\boldsymbol{x} \in \mathcal{S}_{\boldsymbol{x}}$ onto feature patterns $\boldsymbol{y} \in \mathcal{S}_{\boldsymbol{y}}$. Thereby, the function Φ should support the subsequent layer (the adaptive NGnet) in performing the classification task. We consequently strive for an extraction of class-discriminative features that are suited to distinguish between the members and the non-members of a word meaning category. As illustrated in Fig. 4.9, the feature extraction layer comprises two stages. The first stage generates a discriminative feature space, whereas the second stage diagonalizes its dimensions and performs a dimensionality reduction.

Maximizing Renyi's Mutual Information (MRMI)

To generate a class-discriminative feature space, an approach called *Maximizing Renyi's Mutual Information (MRMI)* (Hild et al., 2006) is used. This technique relies on the information-theoretic criterion of the mutual information

$$I(\boldsymbol{Y}; C) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|C) \tag{4.35}$$

between the feature patterns and their category memberships. Here, $H(\boldsymbol{Y})$ and $H(\boldsymbol{Y}|C)$ denote the entropy and the conditional entropy according to Shannon. The mutual information $I(\boldsymbol{Y}; C)$ describes the amount of information that the feature patterns carry about the category memberships. It is hence an appropriate measure for the quality of the feature extraction layer. Since an increase in the mutual information signals an improved discriminability between category members and non-members, the transformation Φ should result in feature patterns that maximize $I(\boldsymbol{Y}; C)$.

Hild et al. (2006) proposed an efficient implementation of the mutual information criterion. It relies on the use of Renyi's quadratic entropy $H_2(\boldsymbol{Y})$ (instead of Shannon's) and its estimation using Parzen windows. By doing so, the mutual information can be approximated on the basis of individual samples. Therefore, let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a set of input samples and $\boldsymbol{Y} = \Phi(\boldsymbol{X}) = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ the corresponding set of feature patterns. Furthermore, let $C = \{c_1, \ldots, c_N\}$ denote the associated class memberships with $c_k \in \{-1, +1\}$. Then the mutual information is calculated as

$$\begin{aligned} I_2(\boldsymbol{Y}; C) &= H_2(\boldsymbol{Y}) - H_2(\boldsymbol{Y}|C) \\ &= -\log\left(\frac{1}{N}\sum_{k=1}^{N} G(\boldsymbol{y}_k - \boldsymbol{y}_{k+1}, 2\sigma\boldsymbol{I})\right) \\ &\quad + \sum_{j\in\{-1,+1\}} \frac{N_j}{N} \cdot \log\left(\frac{1}{N_j}\sum_{k=1}^{N_j} G(\boldsymbol{y}_k^{(j)} - \boldsymbol{y}_{k+1}^{(j)}, 2\sigma\boldsymbol{I})\right). \end{aligned} \tag{4.36}$$

Thereby, $G(\boldsymbol{y}, \boldsymbol{\Sigma}) = \exp(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}))$ is a Gaussian kernel, $\boldsymbol{y}_k^{(+1)}$ and $\boldsymbol{y}_k^{(-1)}$ are feature patterns corresponding to category members and non-members, respectively, N_{+1} and N_{-1} are the numbers of such patterns, and $N = N_{+1} + N_{-1}$ is the overall size of the training set.

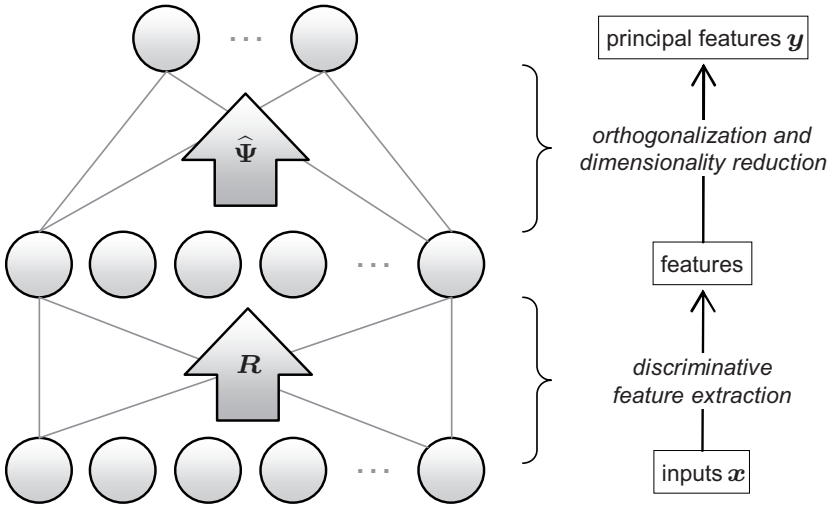


Figure 4.9.: The architecture of the feature extraction layer.

The mutual information criterion can be used to learn a feature extraction function. Here, we restrict learning to a linear feature extraction of form $\boldsymbol{y} = \boldsymbol{R} \cdot \boldsymbol{x}$. We consequently aim at the identification of a transformation matrix $\boldsymbol{R}$, such that the mutual information between the feature patterns and the class memberships is maximized. This can be achieved via stochastic gradient ascent on $I_2(\boldsymbol{Y}; C)$. More precisely, $\boldsymbol{R}$ is iteratively updated according to

$$\boldsymbol{R}_t = \boldsymbol{R}_{t-1} + \eta_{MRMI} \cdot \frac{\partial I_2(\boldsymbol{Y}; C)}{\partial \boldsymbol{R}_{t-1}}, \tag{4.37}$$

where η is a learning rate. A detailed description of the method of Hild et al. (2006) as well as a derivation of the important formulas is given in Appendix B.

Principal Feature Component Space

The MRMI method produces a discriminative feature space. However, it does not indicate which feature dimensions are most important for the representation of a word meaning or how many feature dimensions are needed at all. To answer these questions *Principal Component Analysis (PCA)* is subsequently applied on the extracted features. This serves a de-correlation of the feature dimensions as well as a dimensionality reduction. PCA boils down to solving the eigenproblem for the covariance matrix of the feature patterns. Thereby, the covariance is given by

$$\begin{aligned}
\operatorname{cov}(\boldsymbol{Y}, \boldsymbol{Y}) &= E\left[(\boldsymbol{Y} - E[\boldsymbol{Y}]) \cdot (\boldsymbol{Y} - E[\boldsymbol{Y}])^T\right] \\
&= E\left[(\boldsymbol{R} \cdot \boldsymbol{X} - E[\boldsymbol{R} \cdot \boldsymbol{X}]) \cdot (\boldsymbol{R} \cdot \boldsymbol{X} - E[\boldsymbol{R} \cdot \boldsymbol{X}])^T\right] \\
&= E\left[\boldsymbol{R} \cdot (\boldsymbol{X} - E[\boldsymbol{X}]) \cdot (\boldsymbol{X} - E[\boldsymbol{X}])^T \cdot \boldsymbol{R}^T\right] \\
&= \boldsymbol{R} \cdot E\left[(\boldsymbol{X} - E[\boldsymbol{X}]) \cdot (\boldsymbol{X} - E[\boldsymbol{X}])^T\right] \cdot \boldsymbol{R}^T \\
&= \boldsymbol{R} \cdot \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X}) \cdot \boldsymbol{R}^T,
\end{aligned} \tag{4.38}$$

where E denotes the expectation operator.

W.l.o.g. the input patterns $\boldsymbol{X}$ are assumed to be white with zero mean and unit variance, i.e. $\mathrm{cov}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{I}$. The principal components consequently can be obtained via eigendecomposition of $\boldsymbol{R} \cdot \boldsymbol{R}^T$. Let $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_K]$ be the resulting eigenvectors and $\boldsymbol{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ the associated eigenvalues. Then the principal component feature space is calculated by

$$\boldsymbol{Y} = (\boldsymbol{\Psi}^T \cdot \boldsymbol{R}) \cdot \boldsymbol{X}. \tag{4.39}$$

The eigenvalues $\boldsymbol{\Lambda}$ represent the distribution of the features' energy among each of the principal components. Consequently, one can restrict feature extraction to those dimensions whose cumulative energy content exceeds a pre-defined threshold θ_{PCA} with $0 \leq \theta_{PCA} \leq 1$ (e.g. $\theta_{PCA} = 95\ \%$). Therefore, let the columns of $\boldsymbol{\Psi}$ be arranged such that their associated eigenvalues are sorted in descending order, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$, and let $\mathcal{E}(l)$ be the cumulative energy content among the first l principle feature components, i.e. $\mathcal{E}(l) = \sum_{i=1}^{l} \lambda_i / \sum_{j=1}^{K} \lambda_j$. Then we choose $\widehat{\boldsymbol{\Psi}}$ according to

$$\widehat{\boldsymbol{\Psi}} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{\mathcal{D}_y}] \quad \text{with} \quad \mathcal{E}(\mathcal{D}_y - 1) < \theta_{PCA} \leq \mathcal{E}(\mathcal{D}_y). \tag{4.40}$$

Taking into account that $\widehat{\boldsymbol{\Psi}}$ forms an orthonormal basis with $\widehat{\boldsymbol{\Psi}}^{-1} = \widehat{\boldsymbol{\Psi}}^T$, the feature extraction stage can be summarized by its feature extraction function

$$\boldsymbol{Y} = \Phi(\boldsymbol{X}) = (\widehat{\boldsymbol{\Psi}}^T \cdot \boldsymbol{R}) \cdot \boldsymbol{X} \tag{4.41}$$

as well as its inverse transformation

$$\boldsymbol{X} = \Phi^{-1}(\boldsymbol{Y}) = (\boldsymbol{R}^{-1} \cdot \widehat{\boldsymbol{\Psi}}) \cdot \boldsymbol{Y}. \tag{4.42}$$

4.2.4. Putting the Pieces together

The simultaneous extraction of word-specific features and learning of word meaning categories poses several problems to a computational model. Firstly, mutual information based feature extraction is a statistical learning method. It presupposes a large training corpus in order to reliably estimate the necessary probabilities on a per sample basis. During online learning, however, samples sequentially arise. In the present framework this problem is circumvented as suggested by CLS theory. The adaptive NGnet serves as a training sample generator, insofar as it reactivates memorized associations based on its internal category representations. These samples finally serve as a training set for the feature extraction layer. Secondly, the feature extraction layer produces the feature space that the categorization layer is operating on. This means that the NGnet has to cope with a continuously changing feature space, since learning alters the space during online operation. This includes a change in the number of feature dimensions as well as a change in the individual dimensions. How to adapt an NGnet to an altered feature space without the need for re-training hence constitutes a fundamental question that has to be answered. A detailed description of the solutions employed by our computational model is given in the following.

Reactivation of Memorized Associations

An NGnet is a generative model for universal function approximation. In other words, in addition to calculating the output $\widetilde{\boldsymbol{c}}(\boldsymbol{y})$ of supplied feature patterns $\boldsymbol{y}$, it provides means to generate samples $\boldsymbol{y}$ given an network output $\widetilde{\boldsymbol{c}}$. With respect to word meaning acquisition, the adaptive NGnet hence allows to generate a set of samples $(\boldsymbol{y}', \boldsymbol{c}')$ that comprises feature patterns $\boldsymbol{y}'$ corresponding to category members ($c' = +1$) as well as non-members ($c' = -1$). Such a generative process can be formally done by drawing K samples $\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_K$ from the distribution $p(\boldsymbol{y}'|\boldsymbol{c}' = \widetilde{\boldsymbol{c}}, \boldsymbol{\Theta})$ with $\widetilde{\boldsymbol{c}}$ either being $+1$ or -1.

To do so, we first determine the posterior probability that the i-th expert produces an output $\widetilde{\boldsymbol{c}}$. Using Bayes' rule this probability can be calculated as follows

$$\begin{aligned} p(i|\boldsymbol{c} = \widetilde{\boldsymbol{c}}, \boldsymbol{\Theta}) &= \frac{p(\boldsymbol{c} = \widetilde{\boldsymbol{c}}|i, \boldsymbol{\Theta}) \cdot p(i)}{\sum_{j=1}^{M} p(\boldsymbol{c} = \widetilde{\boldsymbol{c}}|j, \boldsymbol{\Theta}) \cdot p(j)} \\ &= \frac{G(\widetilde{\boldsymbol{c}}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i) \cdot \rho_i}{\sum_{j=1}^{M} G(\widetilde{\boldsymbol{c}}, \boldsymbol{\alpha}_j, \boldsymbol{\Gamma}_j) \cdot \rho_j}. \end{aligned} \tag{4.43}$$

Thereby, ρ_i denotes the importance of the i-th expert. We consequently determine the number of samples that each expert should generate by drawing K samples from $p(i|\boldsymbol{c} = \widetilde{\boldsymbol{c}}, \boldsymbol{\Theta})$. Let $K_1, \ldots, K_M$ be the result of this process.

What remains is to draw K_i samples from the distribution $p(\boldsymbol{y}|i, \boldsymbol{\Theta}) = G(\boldsymbol{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for each expert i. Therefore, let $\boldsymbol{Z}_i = [\boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,K_i}]$ with $\boldsymbol{z}_{i,j} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then the samples $\boldsymbol{Y}'_i = [\boldsymbol{y}'_{i,1}, \ldots, \boldsymbol{y}'_{i,K_i}]$ can be calculated by

$$\boldsymbol{Y}'_i = \boldsymbol{\mu}_i + \boldsymbol{A} \cdot \boldsymbol{Z}_i, \tag{4.44}$$

where $\boldsymbol{A}$ is obtained from the Cholesky decomposition $\boldsymbol{A} \cdot \boldsymbol{A}^T = \boldsymbol{\Sigma}_i$. An illustration of this process is given in Fig. 4.10. There, (a) shows an example of a binary classification task as well as the hidden units, that the adaptive NGnet uses to solve this task. In (b) the reactivated associations are shown. As can be seen, the model does not reactivate the exact memorized associations (given by the centers of the Gaussian receptive fields), but slight variations of it. This allows the network to generate as many samples as necessary even though just a restricted set of observations may have been memorized.

The transformation of the generated feature patterns $\boldsymbol{y}'$ via the inverse of the feature extraction matrix Φ, i.e.

$$\boldsymbol{x}' = \Phi^{-1}(\boldsymbol{y}') = (\boldsymbol{R}^{-1} \cdot \widehat{\boldsymbol{\Psi}}) \cdot \boldsymbol{y}', \tag{4.45}$$

finally results in a set of samples $(\boldsymbol{x}', c')$. This set can be used to train the feature extraction layer, i.e. for discovering those dimensions that best discriminate the members from the non-members of the word category.

Adaptation to Changed Feature Space

Since the training of the feature extraction layer continuously changes the produced feature space, the internal representation of a category has to be adapted to it. An

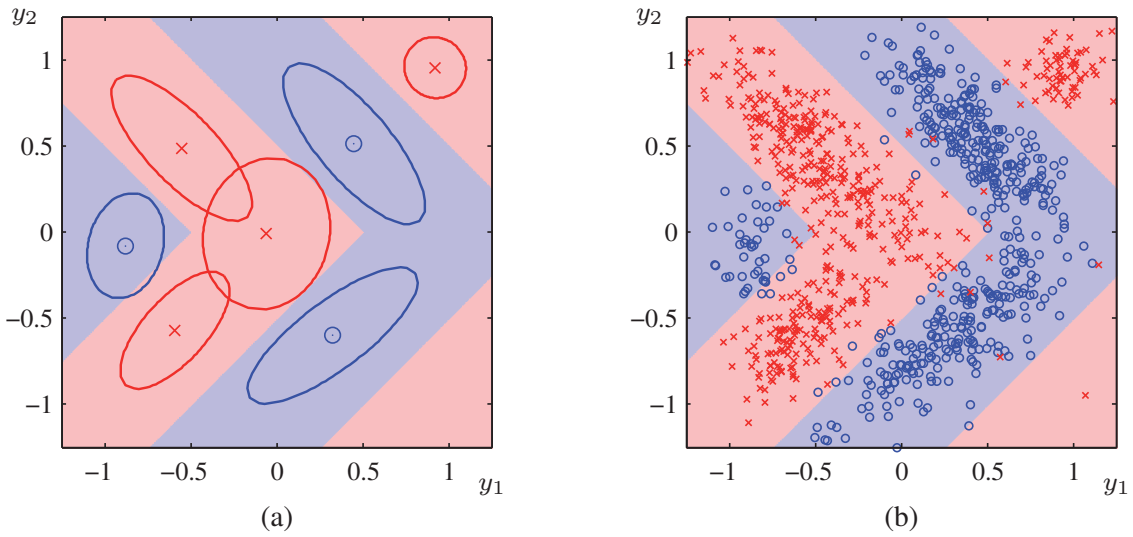


Figure 4.10.: Illustration of the reactivation of memorized associations in the 2-dimensional binary classification task $c = \text{sign}(\cos(\max(y_1 - y_2, y_1 + y_2)))$: (a) shows the receptive fields of the NGnet's hidden units and (b) depicts the samples $\boldsymbol{y}'$ that are reactivated based on them. Blue circles and red crosses correspond to samples with $c' = -1$ and $c' = +1$, respectively.

obvious way to do so is to re-train the NGnet every time the feature space changes. This is of course a time consuming process and further necessitates a memorization of all training samples. Another way is to project the receptive fields of the NGnet's hidden units into the new feature space. For a linear feature extraction like the one proposed in Section 4.2.3 this projection fortunately can be described by a linear transformation matrix $\boldsymbol{B}$ with

$$\begin{aligned} \boldsymbol{B} &= \left(\widehat{\boldsymbol{\Psi}}_t^T \cdot \boldsymbol{R}_t\right) \cdot \left(\widehat{\boldsymbol{\Psi}}_{t-1}^T \cdot \boldsymbol{R}_{t-1}\right)^{-1} \\ &= \widehat{\boldsymbol{\Psi}}_t^T \cdot \boldsymbol{R}_t \cdot \boldsymbol{R}_{t-1}^{-1} \cdot \widehat{\boldsymbol{\Psi}}_{t-1}. \end{aligned} \tag{4.46}$$

Here, $\widehat{\boldsymbol{\Psi}}_{t-1}$ and $\boldsymbol{R}_{t-1}$ as well as $\widehat{\boldsymbol{\Psi}}_t$ and $\boldsymbol{R}_t$ denote the feature extraction matrices before and after a learning step at time t. Consequently, the i-th expert's Gaussian receptive field is adapted to

$$p(\boldsymbol{y}|i, \boldsymbol{\Theta}) = G(\boldsymbol{y}, \boldsymbol{B} \cdot \boldsymbol{\mu}_i, \boldsymbol{B} \cdot \boldsymbol{\Sigma}_i \cdot \boldsymbol{B}^T). \tag{4.47}$$

Improving Generalization

The generalization capabilities of the overall system can be improved in several ways. One has been already mentioned in Section 4.2.3, that is the pruning of feature dimensions based on their relevance for the classification task. Moreover, it is possible to incorporate the relevance of feature dimensions into the process of merging local experts. The underlying idea is as follows: Local experts which are separated along important feature dimensions should stay separated since this difference covers important aspects of the classification task at hand. In contrary, the merging should be preferably done along unimportant feature dimensions. This can be achieved by incorporating the relevance of

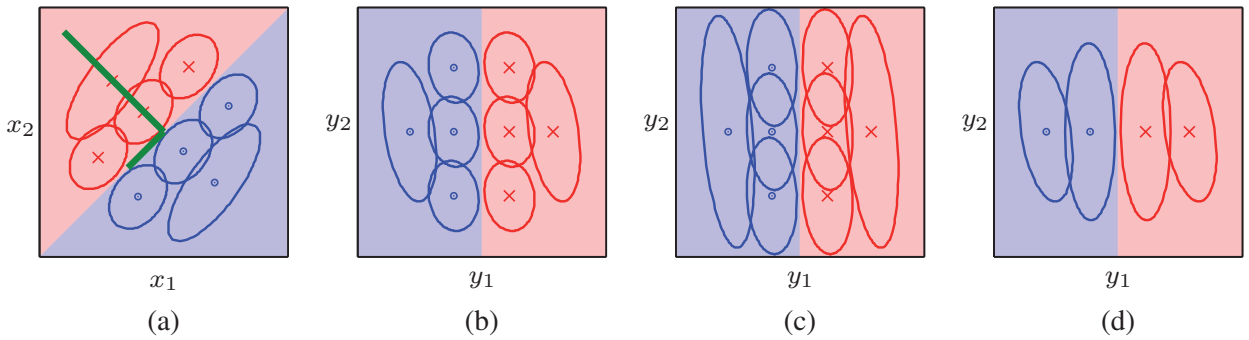


Figure 4.11.: The improved generalization procedure is illustrated on the example of the binary classification task $c = \text{sign}(x_2 - x_1)$. Training the NGnet in this task may result in local experts which partition the input space as shown in (a). Based on these prototypical associations, the feature extraction layer can extract the principal feature dimensions (green bars) in conjunction with their relevance (bar lengths). The NGnet is adapted by projecting the hidden units into the new feature space as shown in (b). The feature importances are used during the merging of the hidden units. Thereby, the receptive fields are artificially scaled along unimportant dimensions as shown in (c), which results in an increased overlap between them. Based on the hidden unit similarity, the experts finally become merged which results in the hidden unit layout depicted in (d).

feature dimensions into the calculation of the similarity between local models. Therefore, let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of the principle feature dimensions (see Section 4.2.3). Then we construct a transformation matrix $\boldsymbol{W}$ as follows

$$\boldsymbol{W} = \begin{pmatrix} \sqrt{\lambda_{max}/\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_{max}/\lambda_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{\lambda_{max}/\lambda_k} \end{pmatrix}, \tag{4.48}$$

where λ_{max} is the maximum eigenvalue, and calculate the similarity between two local models $\mathcal{M}_i$ and $\mathcal{M}_j$ according to

$$\begin{aligned} \mathcal{U}(\mathcal{M}_i, \mathcal{M}_j) \quad = \quad & \mathcal{V}\big(G(\boldsymbol{y}, \boldsymbol{\mu}_i, \boldsymbol{W}\boldsymbol{\Sigma}_i\boldsymbol{W}^T)\,,\, G(\boldsymbol{y}, \boldsymbol{\mu}_j, \boldsymbol{W}\boldsymbol{\Sigma}_j\boldsymbol{W}^T)\big) \\ & \cdot \mathcal{V}\big(G(\boldsymbol{c}, \boldsymbol{\alpha}_i, \boldsymbol{\Gamma}_i)\,,\, G(\boldsymbol{c}, \boldsymbol{\alpha}_j, \boldsymbol{\Gamma}_j)\big). \end{aligned} \tag{4.49}$$

As can be seen, the Gaussians representing the receptive fields of the local experts become scaled, such that they cover larger portions of the input space along unimportant feature dimensions, whereas the coverage along the most important dimension remains unchanged. By doing so, the overlap between the local experts is artificially increased. This finally results in an enhanced merging of hidden units along unimportant feature dimensions and, thereby, yields an improved generalization. Fig. 4.11 illustrates this process on the example of the 2-dimensional binary classification task $c = \text{sign}(x_2 - x_1)$.

Overall Algorithm

Algorithm 4.3 summarizes the proposed method for the acquisition of grounded word meanings. At the beginning of training, the learning system does not have any knowledge on the meaning of a word. More precisely, the feature extraction layer is initialized such that it implements the identity mapping, i.e. feature patterns equal the input patterns. Similarly, the NGnet is empty, i.e. it does not comprise any memorized association between a word and an observation. The subsequent sequential learning scheme allows the network to gain word knowledge based on individual training samples.

The observed word-scene pairs are first memorized by the NGnet. This should lead to an initial increase in the complexity of the network as different hidden units have to be allocated. At the same time, however, a slow feature extraction process searches for commonalities among the memorized associations. Since the feature extraction layer is trained sequentially whenever a new training sample has been observed, knowledge on a word category should gradually shift into the extracted feature dimensions. More precisely, the feature dimensions should become more discriminative with respect to the members and the non-members of a word category. As a result, the system more and more concentrates on the most important aspects of a scene as training progresses (e.g. by pruning less important dimensions). The extracted features should finally represent the *"rules"* that underly the decision whether an input belongs to a word category or not. This eases the classification task and allows the NGnet to generalize the memorized associations. This is done by merging similar hidden units or removing redundant ones. By doing so, the network's size should decrease over time.

The framework consequently implements a gradual transition from an exemplar-based to a rule-based classification system: Initially, the prototypical associations, that are memorized by the NGnet's hidden units, are used to classify new observations in a similarity-based manner. As training progresses, however, the features take over responsibility insofar as they start to represent the rules underlying category membership. The classification of the NGnet therefore becomes context-free as compared to the context-dependent decisions of an exemplar-based classification.

Algorithm 4.3 Algorithm for word meaning acquisition

Initialize the feature extraction matrix to be the identity matrix
Initialize an empty NGnet

for all training samples $(\boldsymbol{x}_t, c_t)$ **do**
  Project the input $\boldsymbol{x}_t$ to the feature space
  Train the NGnet on the new sample $(\boldsymbol{y}_t, c_t)$ using Algorithm 4.2
  Reactivate a set of samples $(\boldsymbol{y}', c')$ comprising category members and non-members
  Project the reactivated feature patterns $\boldsymbol{y}'$ to the input space
  Train the feature extraction layer on the sample set $(\boldsymbol{x}', c')$ using Algorithm B.1
  Adapt the NGnet to the changed feature space
end for

4.2.5. Functional Mapping to Brain Areas

As already mentioned, the computational model is not meant to provide a 1:1 mapping of the system components to specific brain areas. The functional properties of the individual components, however, are largely inspired by neurobiological learning theories. For this reason, a gross mapping with respect to function is reasonable. Here, it is proposed that the computational framework approximately maps onto neurobiological circuits as illustrated in Fig. 4.12.

Firstly, the feature extraction layer is thought to be implemented in neocortical areas. It is known that sensory cortices are organized in form of hierarchies (e.g. the ventral visual pathway $V1 \rightarrow V2 \rightarrow V4 \rightarrow IT$). This hierarchically organized processing transforms low-level input descriptions into higher-level representation. Since such higher-level descriptions are always created for some purpose, they constitute features for tasks. The sensory hierarchies consequently implement a feature extraction and it is reasonable to assume that inputs $\boldsymbol{x}$, features $\boldsymbol{y}$, and words c are represented in neocortical areas.

Finding a homologue to the classification layer is less obvious. On the one hand, the adaptive NGnet can rapidly encode multimodal associations and further is able to reactivate them. These are functions that are typically attributed to the hippocampus. On the other hand, however, the network also employs a statistical learning method that slowly incorporates knowledge. According to Rodríguez-Fornells et al. (2009) a similar dissociation can be found in the MTL structures. More precisely, the authors suggested that the rapid encoding of facts and events primarily relies on the hippocampus and the posterior EC. In contrast, the anterior EC, the PHC, and the PRC may be recruited for a generalization of knowledge via slower learning mechanisms. Evidence in favor of this theory comes from studies with amnesic patients. For example it has been shown that subjects with a brain lesion, that is restricted to the hippocampus, are still able to acquire semantic knowledge – even though a rapid encoding does not seem to be possible anymore. In contrast, patients suffering from a damage of the entire MTL do not show this learning capability (Verfaellie et al., 2000; Bayley and Squire, 2005).

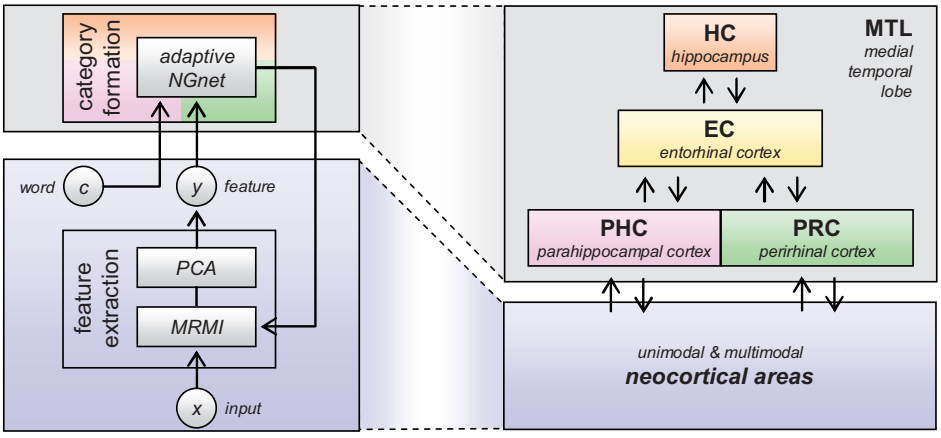


Figure 4.12.: A potential mapping of system components (left) to brain areas (right).

4.3. Evaluation in Benchmarks

The computational model that has been presented in the previous section is not restricted to the domain of word meaning acquisition. It rather comprises learning methods that can be applied in a variety of tasks. For this reason, the framework is first evaluated on standard benchmark problems, which eases a comparison to existing approaches. The benchmarks are taken from the domains of function approximation, binary classification, and categorization. Thereby, the following testing procedure is chosen in order to identify the contributions of the individual system components to the performance of the overall framework: Firstly, the adaptive NGnet is evaluated in isolation, i.e. without the use of the feature extraction layer. Therefore, a function approximation task is used. Secondly, a binary classification problem is used to compare the performance of the NGnet to that of the overall framework, i.e. including the feature extraction layer. Finally, different system configurations in terms of various possible couplings between multiple feature extraction and categorization layers are discussed using a categorization problem.

The performance of the system components is further compared to that of state-of-art approaches. This includes a comparison to the *Resource Allocating Network (RAN)* of Platt (1991) as well as its extensions *RAN-EKF* (Kadirkamanathan and Niranjan, 1993) and *MRAN* (Lu et al., 1997). These methods implement adaptive RBF networks, insofar as new hidden units are allocated based on the novelty of data, and are hence related to the adaptive NGnet presented in Section 4.2.2. The difference between RAN and RAN-EKF is that RAN-EKF updates the network parameters based on extended Kalman filtering instead of the least mean squared algorithm used in RAN. The *Minimum Resource Allocating Network (MRAN)* is similar to the RAN-EKF approach, but further applies a pruning strategy for removing hidden units. Moreover, a comparison to the *Locally Weighted Projection Regression (LWPR)* network by Vijayakumar et al. (2005) is carried out. The LWPR network essentially constitutes an NGnet, but additionally incorporates a local linear feature extraction inside each hidden unit. This allows the network to approximate a target function by locally valid linear models as shown in Fig. 4.13. Besides the concrete learning algorithms employed, the key difference between LWPR and our framework consequently is that LWPR uses multiple local feature spaces as compared to the global feature space of our framework.

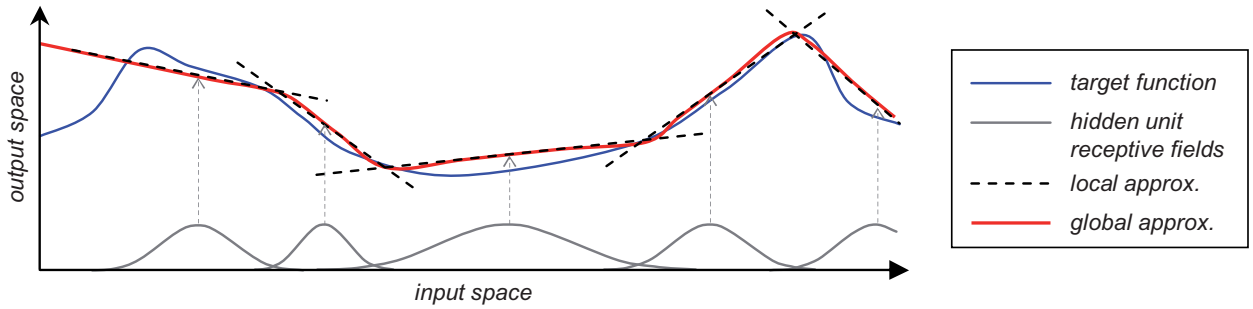


Figure 4.13.: LWPR approximates a target function by overlaying local linear models as compared to the constant approximations employed by an NGnet (cf. Fig. 4.5).

4.3.1. Function Approximation

To evaluate the classification layer, the performance of the adaptive NGnet is assessed in the domain of function approximation. More precisely, the task is the approximation of the two-dimensional cross function which is depicted in Fig. 4.14 and defined as

$$c(\boldsymbol{y}) = \max\left\{ \exp(-10y_1^2),\, \exp(-50y_2^2),\, 1.25 \cdot \exp(-5(y_1^2 + y_2^2)) \right\}. \tag{4.50}$$

As can be seen, the cross function is highly non-linear. This makes it difficult to find an appropriate approximation from only a few training samples. The problem is particularly challenging for incremental learning methods. For this reason, the approximation of the cross function often serves as a benchmark for the comparison of different approaches.

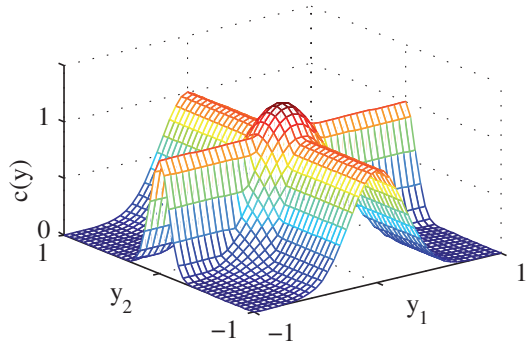
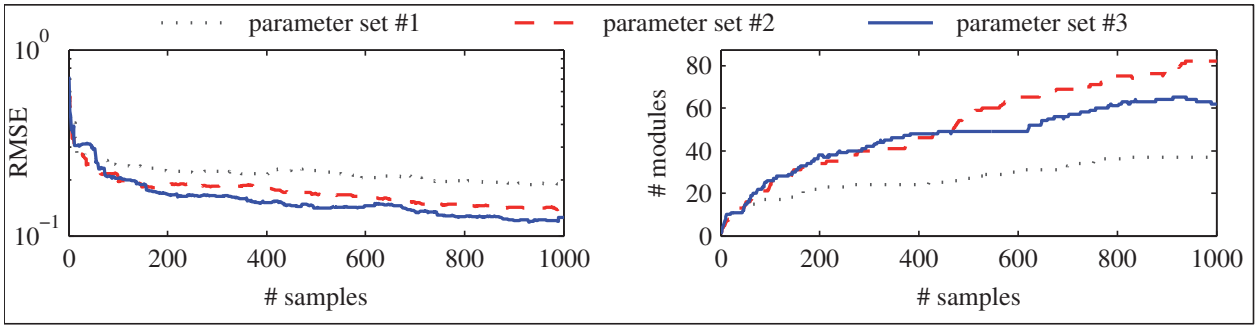


Figure 4.14.: The two-dimensional cross function.

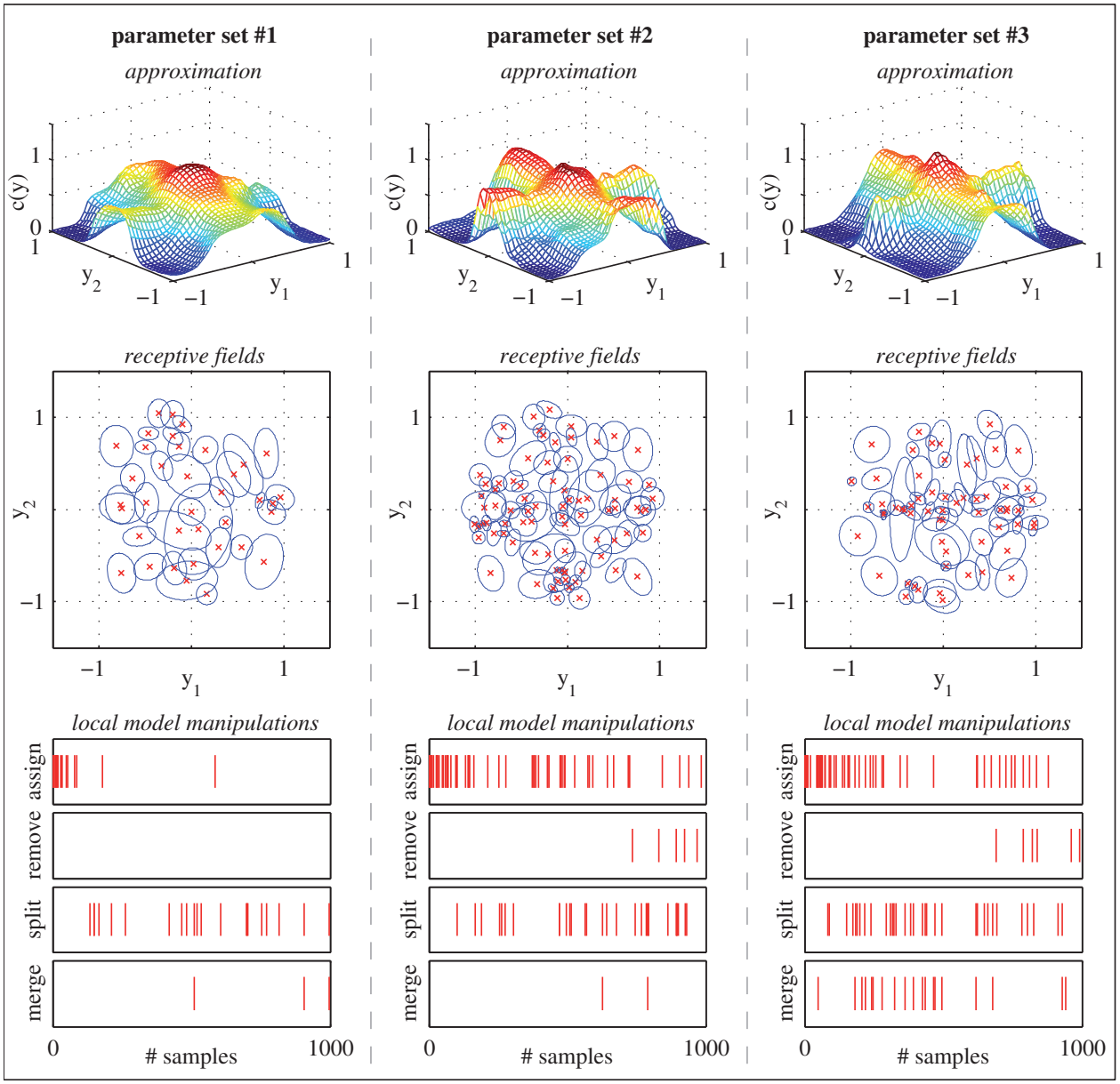
Here, the cross function benchmark is used for the following purposes: Firstly, the adaptive NGnet's learning dynamics are illustrated which includes an investigation of the effect of different parameter settings. Secondly, the NGnet's performance is compared to that of state-of-art networks and, finally, a comprehensive evaluation assesses the noise robustness of the different methods.

Influence of Parameter Settings

During the simulations, training samples $(\boldsymbol{y}, c(\boldsymbol{y}))$ were randomly generated and sequentially presented to the NGnet. The baseline learning rate of the NGnet was always set to $\eta = 0.01$. However, the thresholds for the different model manipulation criteria varied such that their effect on the performance of the NGnet could be estimated. More precisely, we applied the three parameter settings given in Table 4.1. The results for the different simulation runs are depicted in Fig. 4.15. Thereby, (a) shows the evolution of the root mean squared error (RMSE) as well as the number of hidden units as a function of the number of presented training samples. Additionally, for each parameter setting (b) depicts the final approximation of the cross function, the final layout of the hidden units' receptive fields, and the instances in time at which hidden units were assigned, removed, split, or merged.



(a)



(b)

Figure 4.15.: Results for the approximation of the cross function using three different parameter settings: In (a) the evolution of the RMSE and that of the number of hidden units is shown. In (b) more detailed results are depicted, which includes the final approximation, the final layout of the receptive fields, as well as the instances in time when hidden units have been assigned, removed, split, or merged.

	setting #1		setting #2		setting #3
$\theta_{surprise}$	**1.0**	$\longrightarrow$	**0.3**		0.3
$\theta_{coverage}$	**1.0**	$\longrightarrow$	**0.1**		0.1
θ_{remove}	0.01		0.01		0.01
θ_{split}	0.01		0.01		0.01
θ_{merge}	0.8		**0.8**	$\longrightarrow$	**0.6**

Table 4.1.: Different parameter settings.

As can be seen from the results, the adaptive NGnet always approximates the cross function well. Different parameter settings, however, result in varying levels of network complexity and further effect the achieved approximation quality. What all simulation runs have in common is that the RMSE quickly decreases at the beginning. This is shown in Fig. 4.15 (a), where the logarithmic arrangement of the error axis should be noted. In contrast, the number of hidden units quickly increases at the beginning and afterwards converges. This is due to the fact that the network initially allocates many hidden units and thereby grossly approximates the cross function. Subsequently, learning fine tunes the network parameters, which leads to a small increase in the network size and a further decrease in the RMSE. In the following, the ways the different parameters influence the learning are discussed.

Therefore, we first concentrate on the simulation in which parameter set #1 has been applied. From the bottom panels of Fig. 4.15 (b) it can be seen that the network allocates new hidden units almost exclusively at the beginning of training. This is because the thresholds of the criteria for model assignment ($\theta_{surprise}$ and $\theta_{coverage}$) have been chosen very high. More precisely, setting $\theta_{surprise} = 1$ effectively disables unit allocation following large errors at the network output (since the squared approximation error has to be larger than $\theta_{surprise}$). This means that input coverage is the key criterion for unit allocation, which explains the observed unit allocation dynamics: At the beginning of training, units are allocated because many inputs are novel to the network. Later on, however, all inputs are sufficiently covered by already existing units such that no new units are created anymore. As a consequence, the fine tuning of network parameters is mainly carried out via online EM training which does not change the number of hidden units. A small increase in network size is only induced by a splitting of those units that have turned out to provide unreliable approximations. Due to the low number of hidden units, the removal or merging of experts does not play a role in this simulation.

To achieve a finer approximation, the thresholds for unit allocation can be decreased as done in parameter setting #2. By doing so, hidden units are not only assigned at the beginning, but over the whole course of training whenever the network does not approximate the mapping sufficiently well. We consequently obtain a larger network size, but also decrease the RMSE significantly. Due to the fact that the network comprises more hidden units, the number of unreliable experts also increases such that more split events can be observed. The same is true for the removal of units. Since many units are allocated, some of them turn out to be redundant over the course of training. The algorithm subsequently removes such units.

Constructing a small sized network while keeping the approximation quality high, is difficult to achieve. Increasing $\theta_{surprise}$ and $\theta_{coverage}$ and decreasing θ_{merge} as in parameter setting #3 is one way to do this. The latter parameter determines the maximum overlap between hidden units. Decreasing θ_{merge} therefore facilitates small network sizes by generalizing mappings via model merging. However, the parameter have to be chosen with care as too small values can result in overgeneralization. For the current setting, generalization significantly decreases network size and even slightly improves the approximation quality. In summary, the results show that the allocation and merging of hidden units have to be appropriately counterbalanced to achieve the contradictory goals of high approximation qualities and small sized networks. Further simulations (not shown here) revealed just a minor effect of the parameters that control the removal and splitting of units. In fact, largely similar results could be obtained for the cross function approximation task when the parameters were differently chosen.

Comparison to State-of-Art

To judge the performance of the adaptive NGnet, a comparison to existing approaches has been carried out. These methods include *Resource Allocating Networks* (RAN, RAN-EKF, MRAN)[1] and *Locally Weighted Projection Regression* (LWPR)[2]. Furthermore, the cross function has been approximated by *Support Vector Regression* (ϵ–SVR)[3]. Support vector machines constitute powerful nonlinear methods for regression and classification. They are offline trained using a batch of data samples and consequently cannot be directly compared to the aforementioned incremental approaches. SVR should rather be seen as a method against which the performane of incremental methods can be benchmarked.

The results of this comparison are depicted in Fig. 4.16, where the RMSE and the number of hidden units for each network are shown. For ϵ–SVR the number of support vectors is plotted as a homologue of the number of hidden units. As can be seen, RAN and RAN-EKF result in an unconstrained network growth. This is due to the fact that these methods do not include mechanisms for the pruning of hidden units. In contrast, MRAN yields a network size that is approximately the same as that of the adaptive NGnet, but its approximation quality is worse. LWPR finally produces a similar performance as the NGnet and even uses less hidden units. This is because the local linear models of LWPR are more powerful than the constant approximations employed by the NGnet. However, a lot of training data is needed to reliably estimate the linear models which limits LWPR's suitability for rapid learning. LWPR's performance consequently is significantly worse than that of the NGnet at the beginning of training.

[1] The respective networks were implemented according to (Platt, 1991; Kadirkamanathan and Niranjan, 1993; Lu et al., 1997). A number of trials were carried out to determine the parameter setting that yields the best result (learning rate $\nu = 0.01$; growing criteria thresholds $d_{max} = 1.0$, $d_{min} = 0.01$, $d_{decay} = 0.95$, $e_{min} = 0.05$, $e'_{min} = 0.1$; pruning criterion thresholds $M = 25$, $\delta = 0.1$; EKF parameters $P_0 = 1.0$, $R_n = 1.0$, $Q = 0.02$; basis function overlap $\kappa = 0.3$).

[2] The implementation provided at `http://www.ipab.inf.ed.ac.uk/slmc/software/lwpr/` (14.06.2011) was used. The LWPR parameters have been chosen as in (Vijayakumar et al., 2005).

[3] The LIB-SVM implementation of ϵ–SVR with Gaussian kernels was used (Chang and Lin, 2001). A number of trials were carried out to determine the parameter setting that yields the best result ($C = 1000$, $\epsilon = 0.05$, $\sigma = 10$).

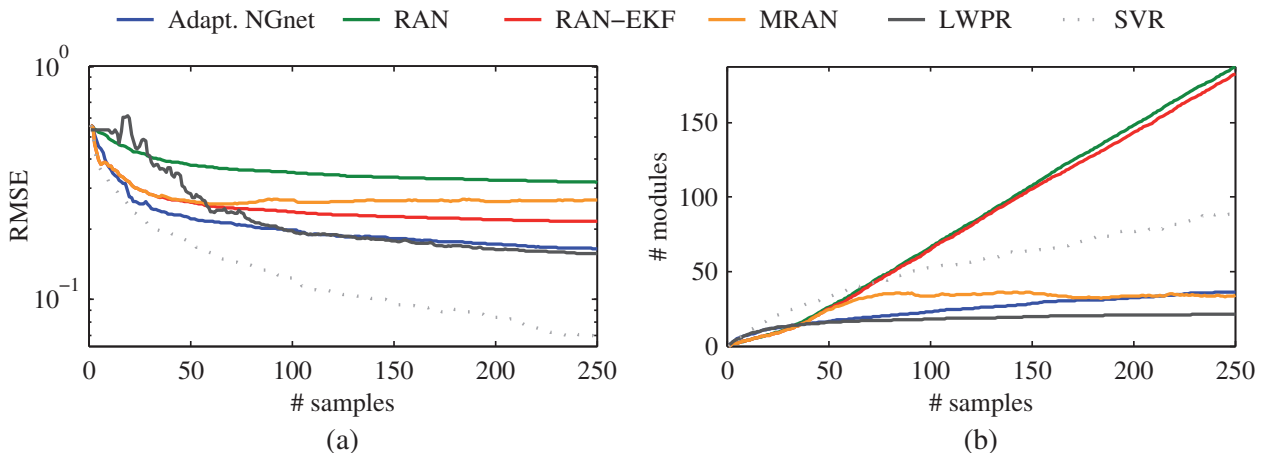


Figure 4.16.: Results for the different methods: (a) the RMSE and (b) the number of hidden units as a function of the number of training samples. The results are averaged over 10 simulation runs.

Noise Robustness

To evaluate the noise robustness of the different methods we added Gaussian noise to the training data. This means that the training samples were generated according to $(\boldsymbol{y}, c(\boldsymbol{y}) + \gamma \cdot n)$ with $n \sim \mathcal{N}(0,1)$. The parameter γ was chosen such that signal-to-noise ratios (SNRs) of $+12 \ldots -6$ dB were obtained. Thereby, an SNR of 0 dB corresponds to equally strong energy levels of the signal and the noise, respectively. A doubling of the signal energy level reflects itself in an increase of the SNR by 3 dB and vice versa. When training the NGnet on the noisy data, we expected a decreased performance and an increased network size when SNR decreases. As shown in Fig. 4.17, this effect can be observed. The network size particularly increases for an SNR of 0 dB or worse. This is due to the fact that the network tries to maintain its approximation quality by recruiting

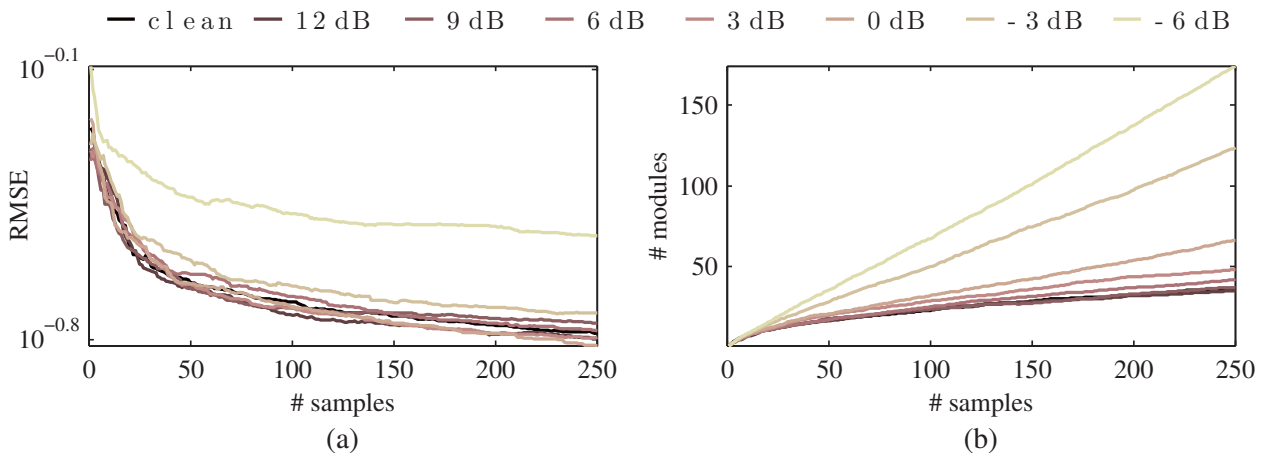


Figure 4.17.: Results of the adaptive NGnet when the training data is contaminated with noise at SNRs of $+12 \ldots -6$ dB. The plots are averaged over 10 simulation runs.

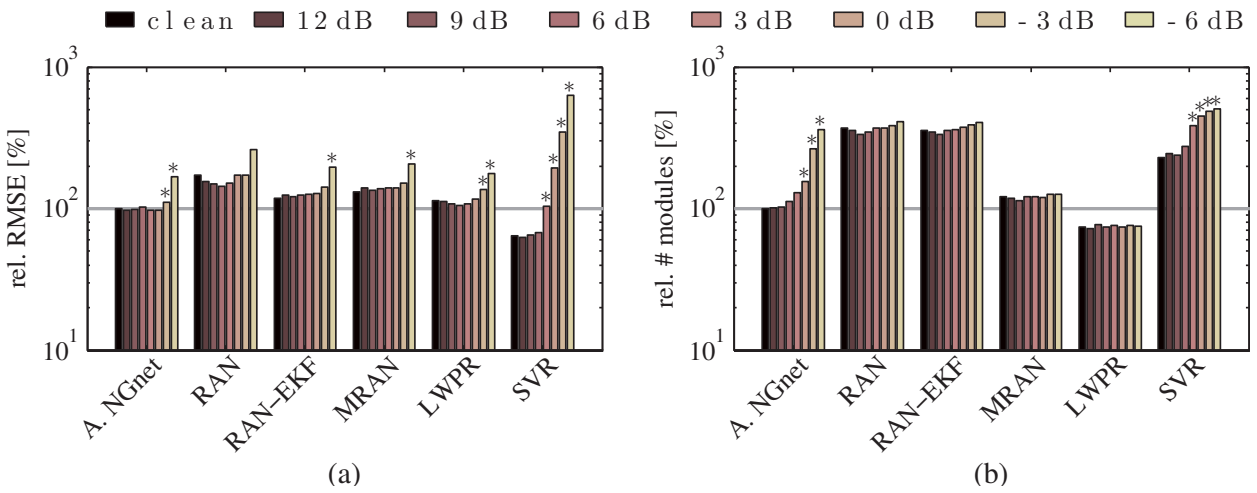


Figure 4.18.: Performance in noise relative to the performance of the adaptive NGnet when a clean training signal is used. Differences within a group of bars can be used to estimate the effect of noise on the performance of the individual methods. Bars marked with "$*$" indicate a setting which yields results that are significantly different to those obtained using clean signals ($p < 0.01$). Significance analysis was based on Welsh's t-test using 10 simulation runs per method and SNR level, where the respective values were averaged over all numbers of training samples.

more hidden units. As a result, just minor performance degradations can be observed for large SNRs. However, the increase in the number of hidden units is not sufficient to counteract performance degradations in case of very small SNRs, e.g. -6 dB.

Similar evaluations were carried out for the state-of-art methods. Fig. 4.18 shows the respective results. There, the RMSE and the number of hidden units for each method and SNR are plotted relative to those of the NGnet using clean training data. By doing so, the plot can be used to investigate whether the adaptive NGnet performs better in noise compared to the other methods. The plot further allows to judge the effect of noise on the individual algorithms. As shown in (a), the RMSE of all methods deteriorates for very small SNR. Thereby, noise particularly affects the SVR results. The NGnet, however, outperforms the other incremental approaches in all cases tested. What distinguishes the NGnet from the other networks is that it tries to compensate the influence of noise by increasing its size. As shown in (b), this is not the case for the other methods. More precisely, no significant influence on network size can be found for RAN, RAN-EKF, MRAN, or LWPR. Only SVR increases the number of support vectors as SNR decreases.

In summary, the evaluations using the cross function benchmark show that the adaptive NGnet achieves a function approximation performance that is comparable to that of existing approaches. It performs slightly better than the resource allocating networks and achieves an approximately similar performance as LWPR. This is true for clean and noisy training data. The evaluations further showed that SVR is strongly affected by noise such that SVR performance gets worse than that of the incremental approaches. Additionally, it is important to note that the RAN and RAN-EKF approaches typically cannot be used due to their unconstrained network growths.

4.3.2. Binary Classification

Next, the computational framework is evaluated in a binary classification benchmark. More precisely, the decision criterion

$$c(\boldsymbol{x}) = \text{sign}(\, x_1 \cdot x_2 - x_3 \cdot x_4 \,) \;\text{ with } x_i \in \mathbb{R}^+ \tag{4.51}$$

is used to assign inputs $\boldsymbol{x}$ to one of two classes, i.e. $c \in \{-1, +1\}$. The reason for using this artificial task is twofold: On the one hand, the classification problem is challenging for incremental learning approaches insofar as it relies on a nonlinear class boundary. On the other hand, however, the task is of limited complexity such that the learning dynamics can be understood and evaluated in detail.

Here, this benchmark is first used to estimate the influence of the different system components on the performance of the overall framework. This is done by comparing the results of the overall framework (classification & simultaneous feature extraction) with those of the NGnet (only classification). Thereby, an investigation of the different learning dynamics reveals the effect that an incorporation of a class-discriminative feature extraction has on the classification layer. As before, the results are compared to those of state-of-art approaches and finally extended by an analysis concerning the noise robustness of the different methods.

Adaptive vs. Static Feature Spaces

To evaluate the model, samples $(\boldsymbol{x}, c(\boldsymbol{x}))$ were randomly generated and finally used to train either the overall framework or only the NGnet. This means that the NGnet was trained on samples $(\boldsymbol{y}, c(\boldsymbol{x}))$ with $\boldsymbol{y} = \Phi(\boldsymbol{x})$. In the first case, the function Φ is obtained by training the feature extraction layer such that Φ continuously changes over time. This means that classification via the NGnet is carried out on an adaptive feature space. In the second case, the function Φ implements the identity mapping, i.e. $\boldsymbol{y} = \Phi(\boldsymbol{x}) = \boldsymbol{x}$, such that classification is carried out on a static feature space. In the simulations, the parameter setup given in Table 4.2 was used. The resulting performances (in terms of the classification error on a test set as well as the number of the NGnet's hidden units) are depicted in Fig. 4.19.

NGnet		Feature Extraction	
η	0.01	η_{MRMI}	0.01
$\theta_{surprise}$	0.4	σ	2
$\theta_{coverage}$	0.1	N	1000
θ_{remove}	0.01	# iter	10
θ_{split}	0.01	θ_{PCA}	0.95
θ_{merge}	0.7		

Table 4.2.: Parameter setting used in the binary classification task.

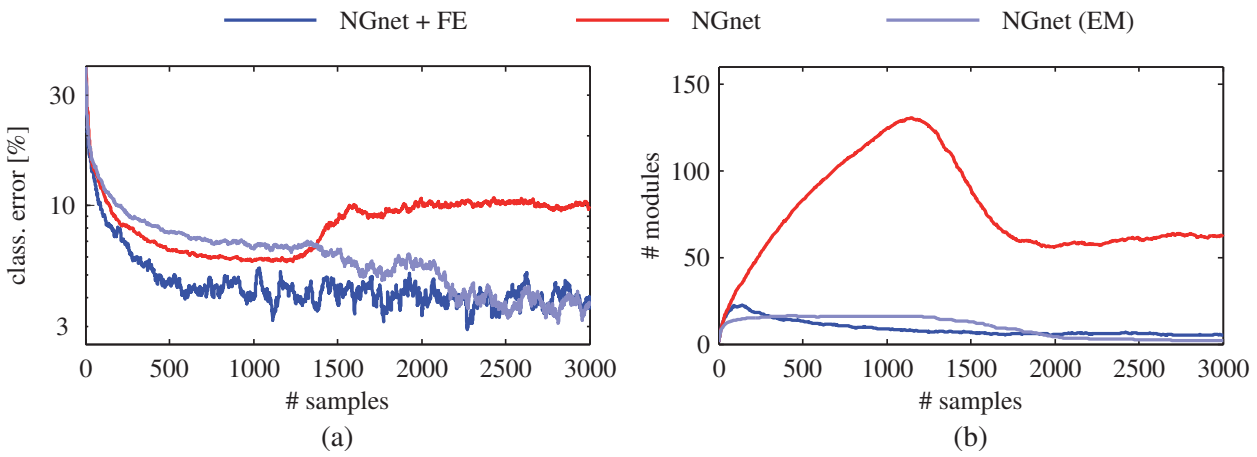


Figure 4.19.: Performance of the model in the binary classification task: (a) depicts the evolution of the classification error, whereas (b) shows the number of hidden units the NGnet is composed of. Thereby, different plots correspond to varying system configurations. This includes a setup comprising an NGnet and a simultaneous feature extraction (NGnet + FE), an independent NGnet (NGnet), and an independent NGnet that solely relies on online EM training (NGnet (EM)).

Here, we first analyze the results obtained by the overall framework (NGnet + FE). As can be seen from the plots, the system is able to acquire class knowledge from few training samples. This reflects itself in the rapid decrease of the classification error at the beginning of training. It can further be observed that the classification layer initially increases its complexity, i.e. the number of hidden units, but starts to decrease it after approximately 100 training samples. The network size subsequently converges to a minimum that is maintained afterwards. One mechanism underlying the observed learning dynamics is the feature extraction, which is analyzed in Fig. 4.20. There, the normalized eigenvalues of the principal feature dimensions are plotted as a function of the number of training samples. The normalized eigenvalues provide a measure for the relative importance of the extracted dimensions. As can be seen, the importance of the different dimensions diverge over time. Whereas one dimension rapidly gains relevance, the other dimensions loose importance or even get pruned. An analysis of the feature extraction matrix Φ

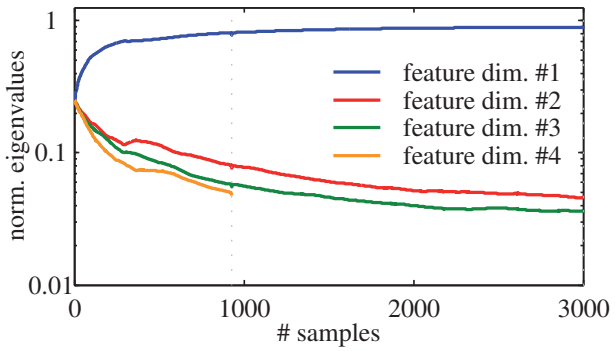


Figure 4.20.: The evolution of the normalized eigenvalues of the principal feature dimensions. The dotted line marks the time when the forth dimension has been pruned.

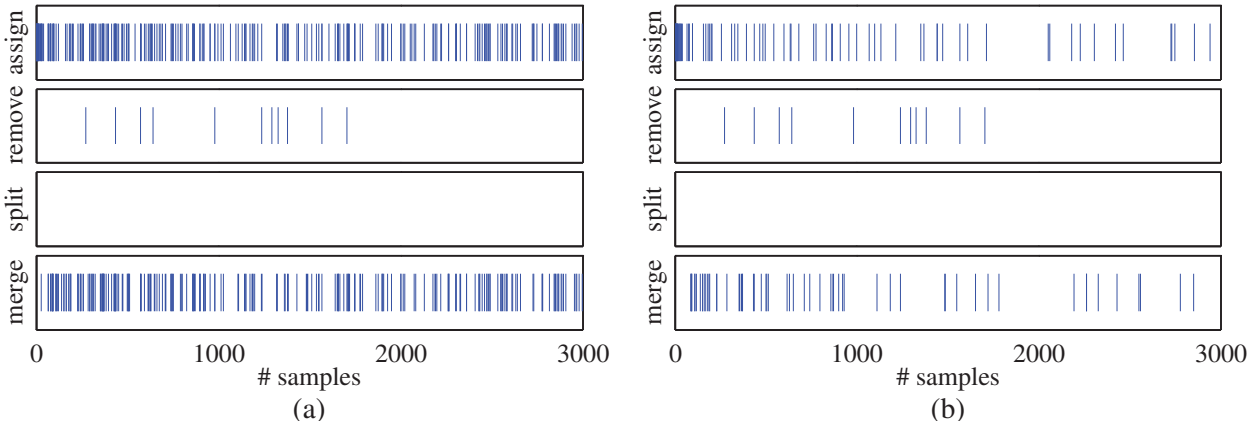


Figure 4.21.: Detailed analysis of the learning dynamics. In (a) the instances in time are depicted where the NGnet assigns, removes, splits, or merges hidden units. In (b) the former plot is corrected, insofar as instances in time where a unit has been assigned and instantaneously merged are not shown.

reveals that the most relevant feature is given by $y_1 \approx (x_1 + x_2) - (x_3 + x_4)$. This feature constitutes an appropriate linear approximation of the real decision criterion (cf. Eq. (4.51)), which demonstrates that the feature extraction layer is able to unveil the relevant class-discriminative information.

To investigate the effect of the feature extraction on the NGnet, Fig. 4.21 (a) depicts details on the NGnet's use of the local model manipulation mechanisms during training. The first aspects that can be observed from the plot is that a splitting of hidden units never occurs. This is because in a classification task, hidden units exclusively represent either members or non-members of a class. Their individual pdfs over the output space $p(c|i, \mathbf{\Theta})$, i.e. over the class labels, consequently degenerates to only one value by which the splitting criterion is never met. This is in contrast to the behavior observed in a function approximation task (cf. Section 4.3.1). The second aspect that becomes evident from the plots is that models are often assigned and merged. Moreover, assignment and merging even seem to coincide. A detailed analysis reveals that this is indeed the case. Hidden units are often assigned following an insufficient prediction of class memberships. Due to the fact that just one feature dimension is of importance, the overlap of newly introduced units with already existing ones is artificially increased along unimportant dimensions by the mechanism described in Section 4.2.4. If this results in a sufficient overlap, the new unit is merged directly after assignment. Existing units consequently incorporate the new observation by merging a model of this observation. Besides the rapid one-shot learning and the slow EM training, this combination of unit assignment and unit merging can be seen as a third type of learning that adapts the network parameters with an intermediate rate.

Fig. 4.21 (b) depicts the plot of (a), where instances of unit assignment and instantaneous merging are excluded. As can be seen, the network mainly allocates units at the beginning of training. Similarly, model merging is particularly observed shortly after the beginning. This coincides with the time, where the feature extraction layer is able to distinguish between relevant and irrelevant feature dimensions. This knowledge facilitates the

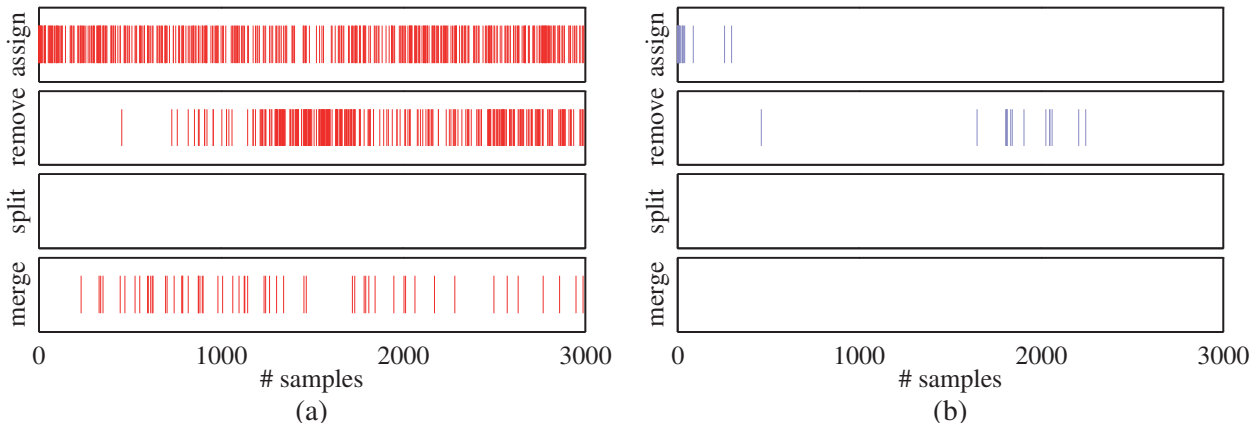


Figure 4.22.: Instances in time when local model manipulation mechanisms have been applied: (a) depicts the results for the normal NGnet, whereas (b) depicts the results for the NGnet that has been forced to rely on EM training.

generalization of the NGnet, insofar as the network can merge hidden units preferably along irrelevant dimensions. This also explains the previously described evolution of the network complexity: At the beginning of training, many observation are novel to the NGnet such that many hidden units are allocated and network size increases. However, as soon as knowledge on the relevance of feature dimensions is acquired, the network uses this knowledge and starts to generalize. This finally yields a minimal network size.

To underpin the importance of a feature extraction, we next analyze the performance of the system that only comprises an NGnet, i.e. no feature extraction layer. As can be seen from the plots of Fig. 4.19, this system results in a much larger network and achieves an inferior performance. Even though a similar kind of pattern concerning the network size can be observed, the generalization phase (where the number of hidden units decreases) seems to negatively effect network performance as the classification error increases. The detailed analysis depicted in Fig. 4.22 (a) reveals that models are assigned over the whole course of training. This is due to the fact that most observations are surprising to the network, because the nonlinear class boundaries prevent the NGnet from a correct generalization to novel situations. As many of the hidden units reflect specific observations, they are of less importance with respect to the performance of the overall network. For this reason, a massive removal of units can be observed later in training. This illustrates that the detection of class-relevant information (in terms of discriminative feature dimensions) is of significant importance for the NGnet.

One could imagine that the NGnet should also be able to extract class-relevant information, since it employs a statistical learning method (EM training). However, it is possible that the network misses to do so by privileging one-shot learning over EM training. To test this hypothesis, a third simulation has been carried out in which the NGnet was forced to apply statistical learning. This was done by increasing the threshold $\theta_{surprise}$ to 1.1 which effectively disabled one-shot learning. As shown in Fig. 4.22 (b), hidden units consequently only become assigned at the very beginning of training, which means that the network parameters are subsequently updated solely based on EM training. As depicted in Fig. 4.19, this indeed enables the network to achieve a network performance

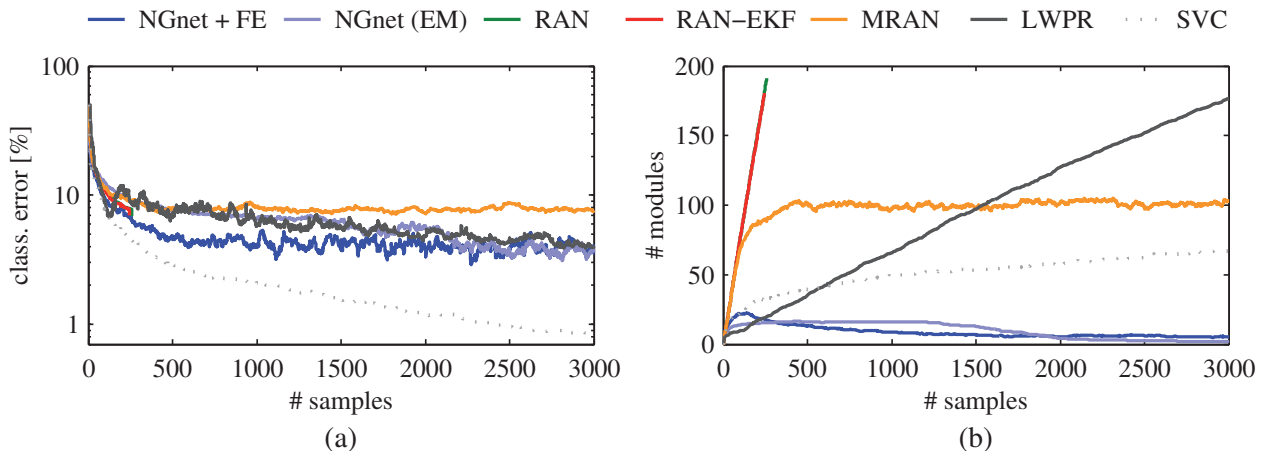


Figure 4.23.: Performance comparison of the different approaches in the binary classification task: (a) depicts the evolution of the classification error, whereas (b) shows the number of hidden units. For SVC, the number of support vectors instead of the number of hidden units is plotted.

that is similar to that of the framework including a feature extraction. However, the NGnet needs much more time to do so, since sufficient amounts of data are needed to reliably exploit the statistics. Overall, this shows that the incorporation of a feature extraction layer enables the framework to rapidly acquire class knowledge while keeping the network size small.

Comparison to State-of-Art

The performance of the computational model has been compared to that of the different RAN[1] networks, LWPR[2], as well as *Support Vector Classification* (SVC)[3]. As before, SVC only served as a benchmark, since it relies on a batch processing of training samples. 10 simulation runs have been carried out for each of the approaches. The averaged results are depicted in Fig. 4.23. It can be seen that RAN and RAN-EKF again yield an unconstrained network growth. For this reason, the respective simulations have been stopped shortly after the beginning of training. MRAN's network sizes converges. However, the approach performs significantly worse than our framework both with respect to classification performance and network complexity. Only LWPR is able to achieve a classification error that is in the range of the one our framework yields. However, learning speed is approximately the same as that of the EM-trained NGnet, i.e. very

[1]The respective networks were implemented according to (Platt, 1991; Kadirkamanathan and Niranjan, 1993; Lu et al., 1997). A number of trials were carried out to determine the parameter setting that yields the best result (learning rate $\nu = 0.01$; growing criteria thresholds $d_{max} = 1.0$, $d_{min} = 0.01$, $d_{decay} = 0.99$, $e_{min} = 0.05$, $e'_{min} = 0.1$; pruning criterion thresholds $M = 25$, $\delta = 0.1$; EKF parameters $P_0 = 1.0$, $R_n = 1.0$, $Q = 0.02$; basis function overlap $\kappa = 0.85$).

[2]The implementation provided at `http://www.ipab.inf.ed.ac.uk/slmc/software/lwpr/` (14.06.2011) was used. The LWPR parameters have been chosen as in (Vijayakumar et al., 2005).

[3]The LIB-SVM implementation of one-class SVM with Gaussian kernels was used (Chang and Lin, 2001). A number of trials were carried out to determine the parameter setting that yields the best result ($C = 10000$, $\sigma = 0.25$).

slow. Moreover, LWPR also leads to an unconstrained network growth. This restricts the application of the approach in more complex scenarios. In summary, our framework comprising an NGnet and a simultaneous feature extraction significantly outperforms the other approaches. This holds with respect to both the rate of knowledge acquisition and particularly the complexity of the employed network.

Noise Robustness

Finally, the noise robustness of the different methods is assessed. Thereby, one can distinguish between noise applied to the inputs and outputs, respectively. In the former case, training samples $(\boldsymbol{x} + \gamma \cdot \boldsymbol{n}, c(\boldsymbol{x}))$ are presented to the system. Thereby, $\boldsymbol{n}$ denotes the noise with $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and factor γ controls the SNR. This means that inputs, which slightly deviate from the true observations, are supplied to the framework. This simulates measurement noise, e.g. noise that is inherent to sensor data or that emerges from inaccurate preprocessing stages. In contrast to this, noise over the output space refers to altering the class memberships of the inputs. This is achieved by using training samples $(\boldsymbol{x}, \widetilde{c}(\boldsymbol{x}))$ where $\widetilde{c}(\boldsymbol{x}) \neq c(\boldsymbol{x})$ for a certain percentage of samples. Output noise consequently refers to wrongly given class labels.

The effects of input noise as well as output noise on the performance of the different approaches are depicted in Fig. 4.24. As can be seen from the plots in (a) and (b), input noise negatively affects the performance of all methods. Thereby, the influence is much more severe than in the function approximation benchmark (cf. Fig. 4.18 where the different scalings of the y-axes should be noted). This is due to the fact that in a classification task, input noise can result in significantly altered outputs – particularly for inputs that are close to class boundaries. We can further observe that the NGnet trained via EM performs slightly better than the framework incorporating a feature extraction in medium to high noise conditions. The reason for this is twofold: Firstly, the feature extraction generates a feature space that is lower-dimensional than the original one. Wrongly assigned hidden units consequently cover larger portions of the input space by which more observations become wrongly classified. Secondly, the excessive use of EM training by the independent NGnet may be beneficial, since statistical learning can average out the noise influence. As shown in Fig. 4.24 (c) and (d), output noise constitutes a much larger problem for the different methods. Even small amounts of wrongly given class labels significantly decrease the performance of all approaches. This is because input noise only sometimes alters the class memberships of observations, whereas output noise always does.

Overall, the evaluation in the binary classification benchmark revealed that both processing modules, the NGnet and the feature extraction, are important for efficient learning. Whereas, an independent application of the NGnet already outperforms existing approaches, the integration of an additional feature extraction further facilitates a rapid learning and generalization. The results demonstrated, that the overall framework constructs small-sized networks that show an excellent classification performance. However, problems with respect to the robustness against wrongly given class labels could be identified (as it is also the case for the other approaches). In Section 4.5 we will discuss possible future improvements that tackle this issue.

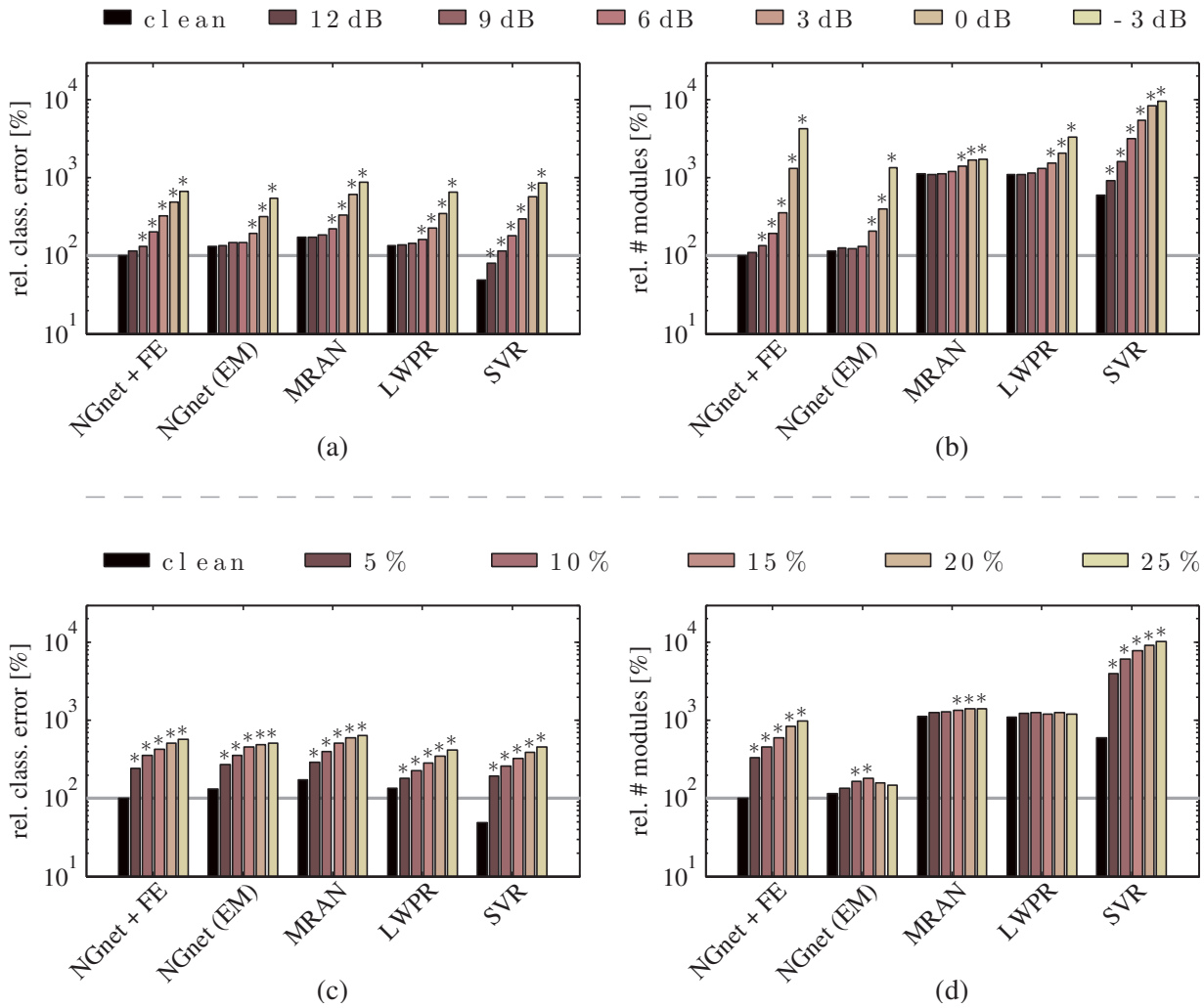


Figure 4.24.: Performance of the different approaches in noise relative to the performance of our framework when a clean training signal is used: (a) and (b) depict the influence of input noise, whereas (c) and (d) show the effect of noise applied to the output. Differences within a group of bars can be used to estimate the effect of noise on the performance of the individual methods. Bars marked with "$*$" thereby indicate a setting which yields results that are significantly different to those obtained using clean signals ($p < 0.01$). Significance analysis was based on Welsh's t-test using 10 simulation runs per method and noise level, where the respective values were averaged over all numbers of training samples.

4.3.3. Categorization

The framework is finally evaluated in the domain of categorization. Thereby, categorization constitutes a multi-class classification task, insofar as it targets the prediction of an observation's membership to multiple classes or categories. Word meanings are a good example in this respect. Each observation typically can be described by multiple words, e.g. the words *Honda*, *car*, and *vehicle* can refer to the same object. In this section, it is discussed how the framework can be used to solve such tasks. A particular emphasis is given on different system configurations. More precisely, we show how multiple NGnets and feature extraction layers can be combined and further investigate the suitability of the resulting configurations.

Global vs. Category-Specific NGnets

As it will be shown in the following, any categorization problem can be transformed into a set of binary classification tasks. Therefore, consider the problem where an observation has to be categorized according to M categories. It is important to note that any observation can belong to multiple categories. This means that we would like to predict the category memberships $\boldsymbol{c}(\boldsymbol{x})$ of an input $\boldsymbol{x}$ with $\boldsymbol{c}(\boldsymbol{x}) = [c_1(\boldsymbol{x}), c_2(\boldsymbol{x}), \ldots, c_M(\boldsymbol{x})]^T$, where $c_i(\boldsymbol{x}) \in \{-1, +1\}$ denotes the input's membership with respect to the i-th category. An M-class categorization problem consequently can be described in terms of M binary classification problems.

Fig. 4.25 shows two different system configuration that can be used to solve such a task. In (a) a feature extraction layer transforms an input $\boldsymbol{x}$ into a feature pattern $\boldsymbol{y}$ which is subsequently categorized using a globally operating NGnet. This means that the NGnet comprises multiple output nodes which each signal the input's membership with respect to one of the categories. The NGnet's hidden units consequently memorize associations between inputs and all categories the inputs belong to. In contrast, the framework depicted in (b) is composed of multiple locally operating NGnets. Thereby, each network is responsible for the representation of a single category, i.e. it comprises a single output node and memorizes associations between inputs and the category of interest.

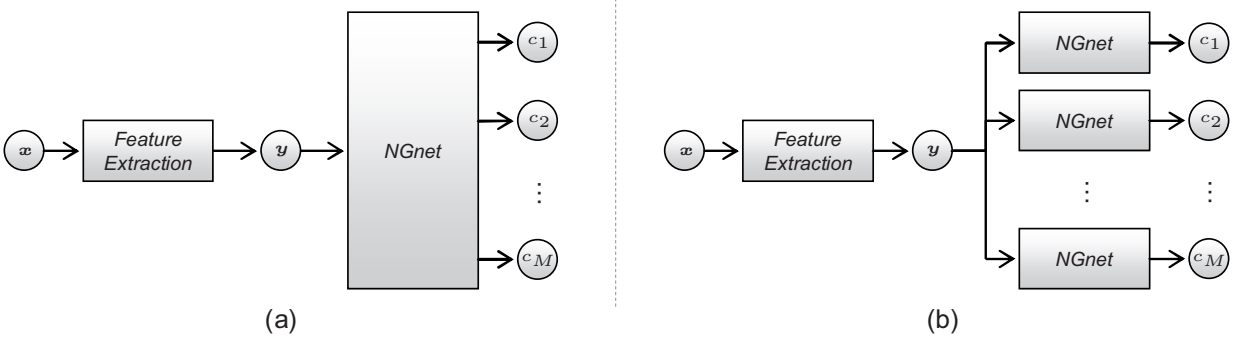


Figure 4.25.: Different system configurations for categorization: In (a) a global NGnet predicts the memberships of an input with respect to all categories, whereas in (b) multiple category-specific NGnets are used to do so.

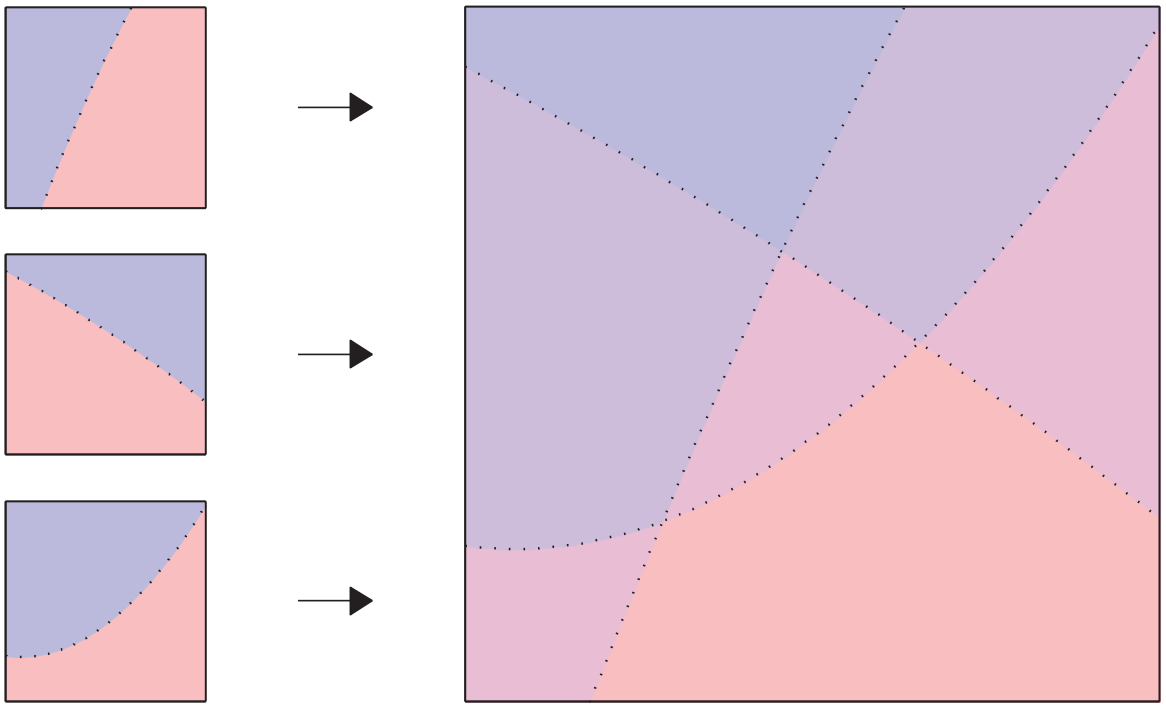


Figure 4.26.: The decision boundaries of three binary classification problems are depicted in (a). Representing the three categories in single network aims at solving the categorization problem depicted in (b). Any combination of category memberships (any closed area in the feature space) has to be represented by an NGnet.

Of course, further schemes can be implemented as combinations of these two extremes. For example, a single NGnet could be used for the representation of K categories, whereas the remaining $M - K$ categories are in the scope of other NGnets.

In the following, the suitability of the different configurations is discussed using the example depicted in Fig. 4.26. In (a) three different categories are shown in terms of their membership decision boundaries in the two-dimensional feature space. By interpreting each category as a binary classification task, each category could be easily represented by a category-specific NGnet. Thereby, the hidden units of the different NGnets need to represent feature patterns as members or non-members of the respective categories, respectively. Since the decision boundaries are simple in this example, this can be done with networks that comprise just a few hidden units. If the same categories should be represented by one global NGnet, the classification problem depicted in (b) has to be solved. More precisely, the NGnet has to cope with any possible combination of individual category memberships (any depicted closed area in the feature space). For example, the network has to comprise hidden units which refer to feature patterns that do not belong to any of the three categories. Similarly, it has to comprise hidden units which cover feature patterns that only belong to the first category, to the second category, to the first and the second category, and so forth. It is obvious that the problem becomes even worse, if more complex decision boundaries or more categories have to be represented. More precisely, the complexity of the NGnet increases exponentially with the number of

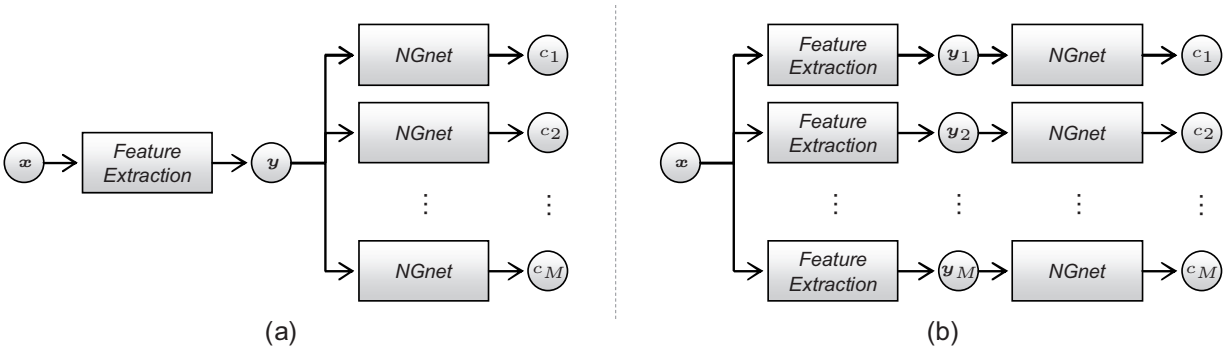


Figure 4.27.: Different system configurations for categorization: The NGnets can either (a) share a global feature space or (b) use individual local feature spaces.

categories. For this reason, the system configuration depicted in Fig. 4.25 (b), i.e. the use of category-specific NGnets, is generally preferable over the application of a global NGnet as depicted in Fig. 4.25 (a). Exceptions to this rule can be found, of course. For example, a single NGnet may be suitable in case of mutually exclusive categories. For such categories the decision boundaries do not overlap, such that the combinatorial explosion does not occur. In the following, however, we will consider the use of category-specific networks.

Global vs. Category-Specific Feature Spaces

Similar to the aforementioned discussion on the use of multiple NGnets, the framework can make use of multiple feature extraction layers. As shown in Fig. 4.27, two configurations are of particular interest: Multiple classification layers (NGnets) can either share a common feature space as depicted in (a) or rely on individual feature spaces as illustrated in (b). In the former case, the aim of the feature extraction layer is to provide a feature space that discriminates the members of the different categories from each other. In other words, feature patterns that belong to one category should become separated from those of other categories. Of course, this often cannot be achieved as a particular feature pattern may belong to multiple categories. This problem is circumvented by the second system configuration. There, a feature extraction layer should produce a feature space in which the members of a particular category should be separable from the non-members of the same category. This means each category is considered independently from the other ones. For this reason, the use of category-specific feature spaces is generally preferable over the use of a global feature space.

For mutually exclusive categories, however, the aforementioned argument does not hold. Nevertheless, it is proposed that category-specific feature spaces are beneficial over a global feature space even in case of non-overlapping categories. The reason for this is as follows. A global feature extraction considers all categories simultaneously. This means, it tries to discriminate each category from all other categories. In contrast, a category-specific feature extraction only tries to discriminate one category from the other categories. This means that the other categories do not have to be distinguishable from each other. A category-specific feature extraction thus constitutes a sub-problem of a global feature

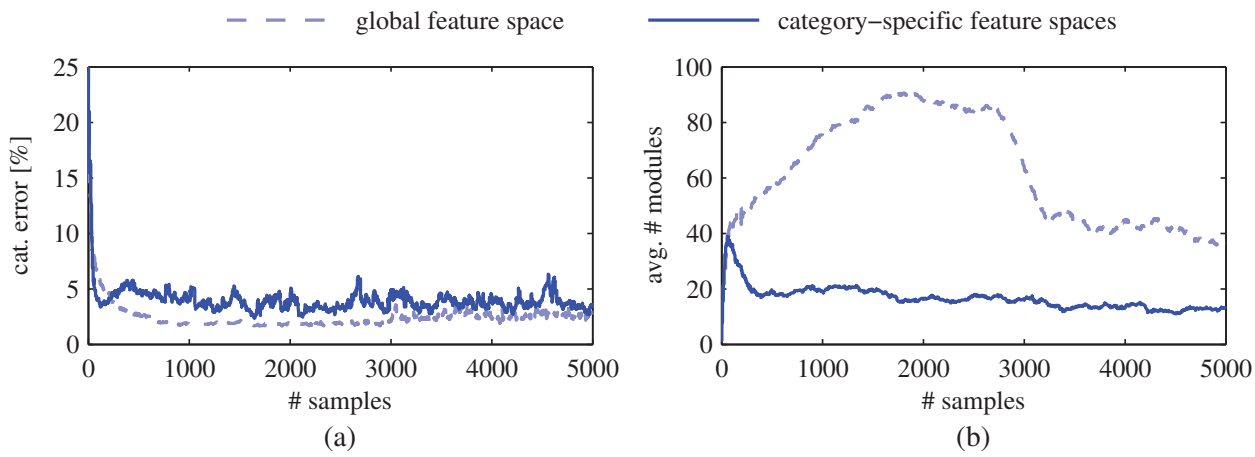


Figure 4.28.: The performance of the systems that either rely on a global feature space or use category-specific feature spaces: (a) shows the categorization errors and (b) the average number of hidden units per NGnet.

extraction and hence should be easier to achieve. Next, this hypothesis is computationally validated in the domain of hand-written letter recognition. The reason for using this task is twofold: Firstly, letters constitute mutually exclusive categories, i.e. each written character can be a member of only one letter category. Secondly, letter recognition is a challenging problem, such that qualitative as well as quantitative differences between the use of both system configurations should become visible.

The evaluation is based on the *Letter Recognition Data Set*[1] of Frey and Slate (1991). It comprises 20000 pixel images of the 26 English capital letters, each of them being represented by 16 integer-valued attributes like their statistical moments or edge counts. For the sake of clarity, in the following we restrict evaluation to every fourth letter, i.e. *A*, *E*, *I*, *M*, *Q*, *U*, and *Y*. The training samples were sequentially presented to two systems, one of them relying on a global feature extraction and the other one using category-specific feature extraction layers. Thereby, we applied the same parameter setting as in the previous experiments (cf. Table 4.2). The categorization performance was calculated on a separate test set which contained 25% of the data samples.

The results of the simulations are depicted in Fig. 4.28, where the evolution of the categorization error as well as the average number of hidden units per NGnet is plotted for both system configurations. As can be seen from the plots, just a minor difference between the two systems can be observed with respect to the achieved categorization errors. In both simulations, the error quickly decreases at the beginning of training and maintains a low level of approximately 3% afterwards. However, the plots show a significant difference in the network complexities. Whereas the global feature space enables the NGnets to generalize after approximately 2500 training samples, generalization occurs much earlier (after ∼100 samples) when using category-specific feature spaces. We further observe that individual feature spaces finally result in less complex NGnets than

[1] Available at the UCI Machine Learning Repository: University of California, School of Information and Computer Science, Irvine, CA, USA, `http://archive.ics.uci.edu/ml/` (14.06.2011).

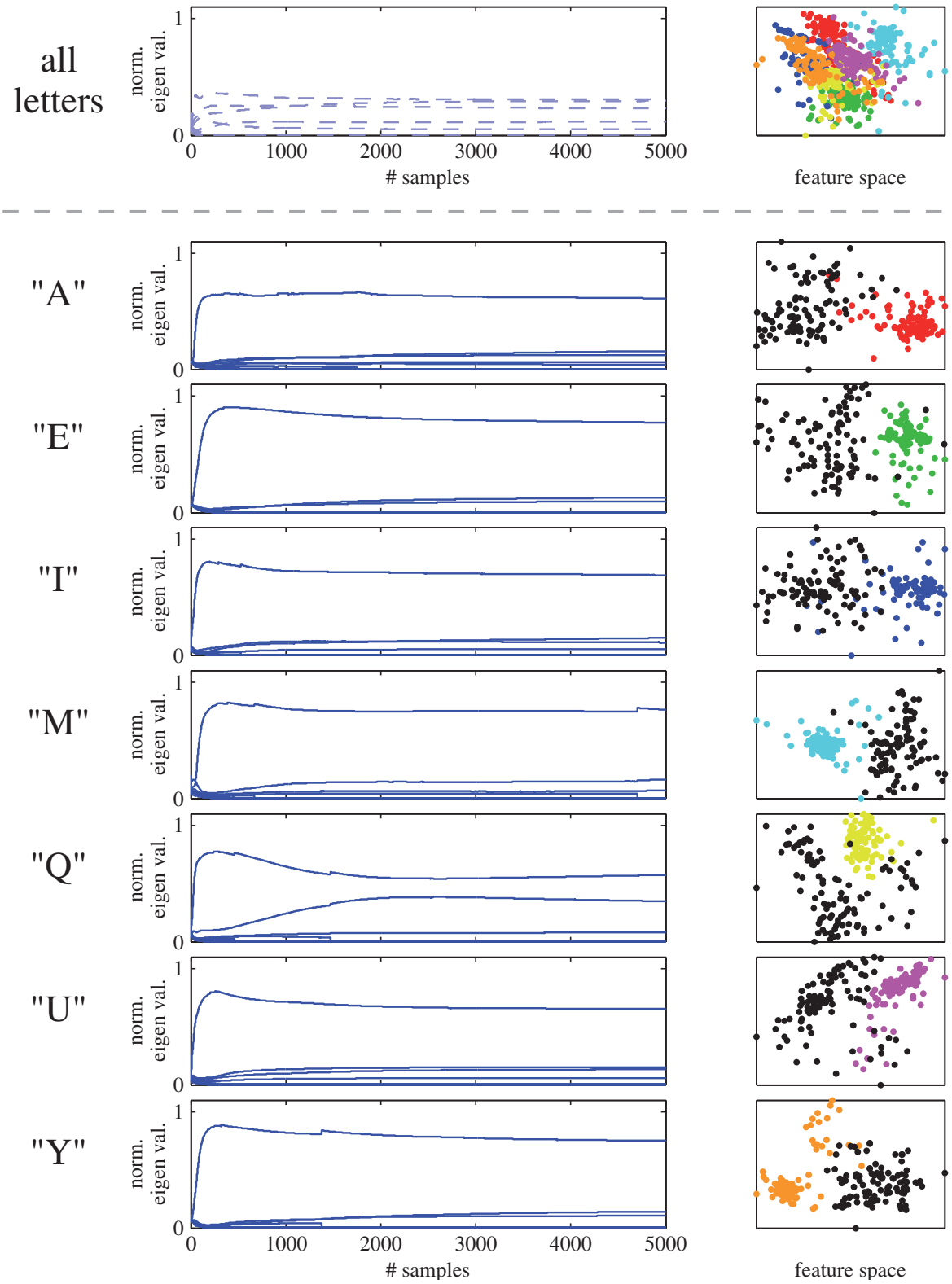


Figure 4.29.: The evolution of the feature dimensions' relevance, i.e. their normalized eigenvalues, for the system using a global feature space (top) and the system using category-specific feature spaces (bottom). The insets at the right show the feature spaces spanned by the first two principal dimensions, respectively. Colored dots correspond to samples of the different categories. For the category-specific feature spaces, black dots refer to non-members of the category (other letters).

a global feature space does. The reason for this behavior is depicted in Fig. 4.29, where the evolution of the normalized eigenvalues of the different principal feature dimensions is plotted. As can be seen, the individual feature extraction layers are able to rapidly extract the relevant feature dimensions. For each letter, processing concentrates on one or two dimensions that are sufficient to discriminate the particular letter from all other categories (see insets at the right). In contrast, the global feature extraction produces a feature space in which more dimensions are relevant. This first becomes evident from the fact that the normalized eigenvalues are less diverse as compared to the category-specific ones. Secondly, the inset at the right shows that the first two principal dimensions are not sufficient to discriminate between the samples of the different letter categories. The global feature space consequently needs more dimensions to distinguish between the letters. These differences in the feature extraction dynamics finally result in the observed differences in the complexities of the two systems. In summary, the experiments suggest that the best results can be obtained by using individual feature extraction layers and NGnets for each category, respectively.

4.4. Application to Word Learning

To demonstrate the model's suitability to acquire word meanings, the framework has been applied in a simulated word learning scenario. The visual scene description task depicted in Fig. 4.30 was used for this purpose. In the scenario a learner and a tutor observe scenes composed of randomly created geometric objects. The task was to learn the meaning of words which describe the relations between the objects. For training the system, the tutor iteratively selects two of the objects, points to them, and simultaneously provides a word label for their relation, e.g. by saying

*"This object **is larger than** that object."*

In the present experiment it is assumed that the learner has knowledge about the grammar of the respective utterances. In other words, the learner knows that the first object is the object of interest, the second object is the reference object, and the middle term (e.g. *is larger than*) denotes the word or phrase to learn. We further assume the learner to possess image processing capabilities that allow him to extract visual properties from the objects. Since the development of such capabilities is out of the scope of this thesis, in the experiment the object properties have been extracted from an internal simulation state. In detail, each object is represented by its absolute center position, its width and height, as well as its RGB color values. A 14-dimensional input vector $\boldsymbol{x}$ (7 dimensions per object) consequently served as a description of a scene. The experiment included word labels for relations concerning the positions of the objects (*is to the left of*, *is to the right of*, *is above*, *is below*), their sizes (*is larger than*, *is smaller than*), and their colors (*is brighter than*, *is darker than*). The model obviously did not have direct access to category-relevant feature dimensions (e.g. the relative object positions); it rather had to extract them autonomously.

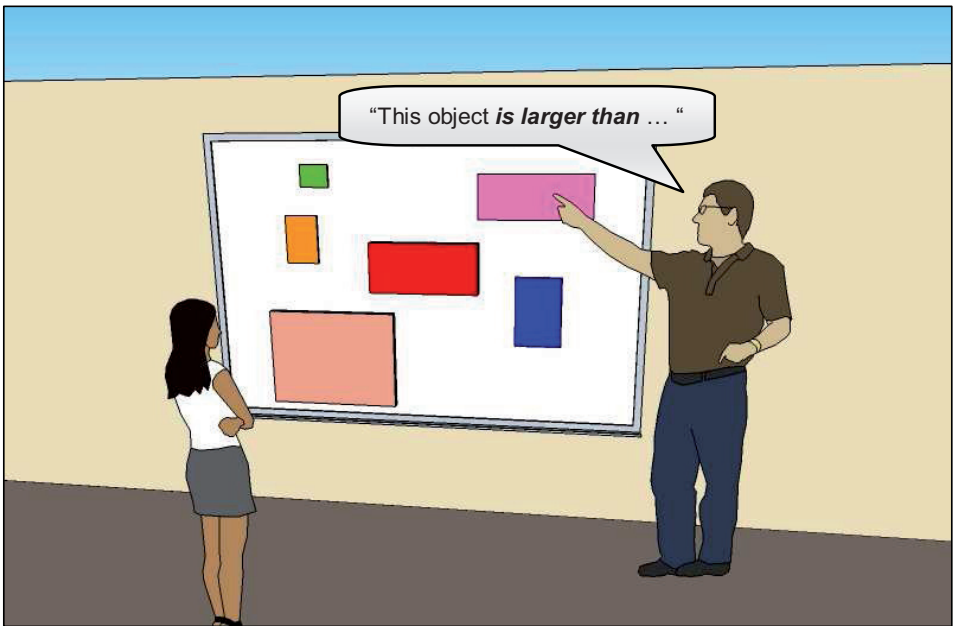


Figure 4.30.: An illustration of the simulated visual scene description task.

To carry out the learning task, 8 instances of the framework have been used, i.e. one categorization and feature extraction layer for each word to learn. These instances have been trained by tuples $(\boldsymbol{x}, \boldsymbol{c})$, where $\boldsymbol{x}$ is the input vector and $\boldsymbol{c} = [c_1, \ldots, c_8]^T$ the vector of category memberships. Here, it is important to note that most c_i are undefined, since just one word label is provided with each sample. This means that $c_i = +1$ only for the category corresponding to the supplied word label. To circumvent the problem of missing negative training data, a mutual exclusivity bias has been implemented. In other words, a positive training exemplar for *is larger than* ($c_i = +1$) has been additionally used as negative sample for *is smaller than* ($c_j = -1$). The bias only has been applied between words related to object positions, sizes, and colors, respectively. The learner consequently has been equipped with innate knowledge on the exclusivity of specific words. Even though children seem to make use of similar learning constraints (Markman and Wachtel, 1988), their innate availability is unrealistic and hence constitutes a restriction of the current experiment. In Chapter 5, however, it will be shown how such a bias can develop over the course of training without any significant change in system performance.

For the evaluation of the model, the correct categorization rate has been calculated on a set of scenes which have not been used for training. The corresponding result curves are depicted in Fig. 4.31, where we additionally plot the number of local experts which comprise the individual NGnets. Fig. 4.31, thus, depicts how (a) system performance and (b) system complexity evolve as a function of the number of training samples. To keep the plots readable, curves for the learning of *is to the left of*, *is larger than*, and *is brighter than* are shown. Similar results have been obtained for the other words. The plots illustrate that the model is able to rapidly acquire the meaning of words. The initially poorer performance for *is to the left of* stems from the fact that more negative than positive training exemplars are used as compared to the learning of the other words. Nevertheless, just a few training samples are needed to achieve a high system performance of ∼80%. This is due to the on-demand allocation of local experts which reflects itself in

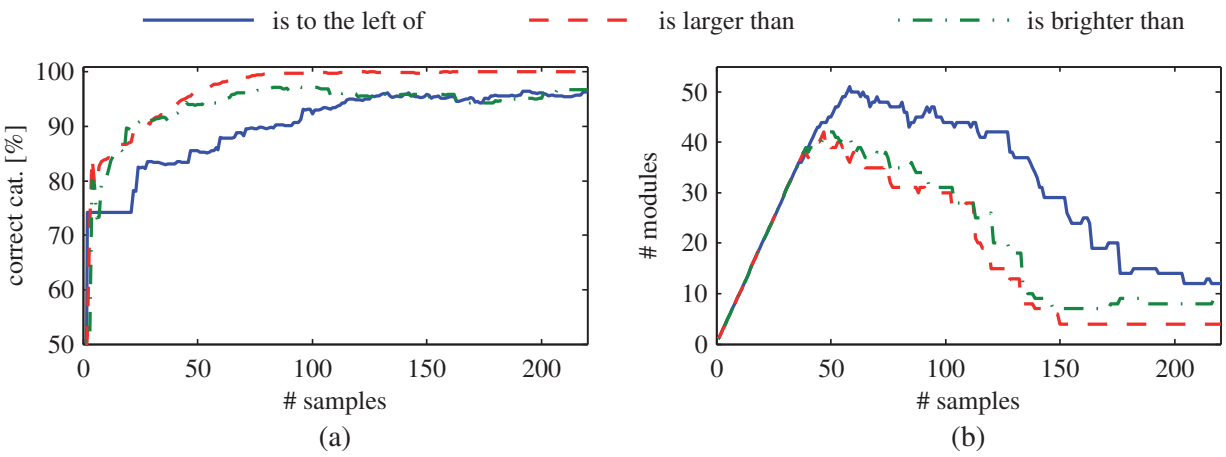
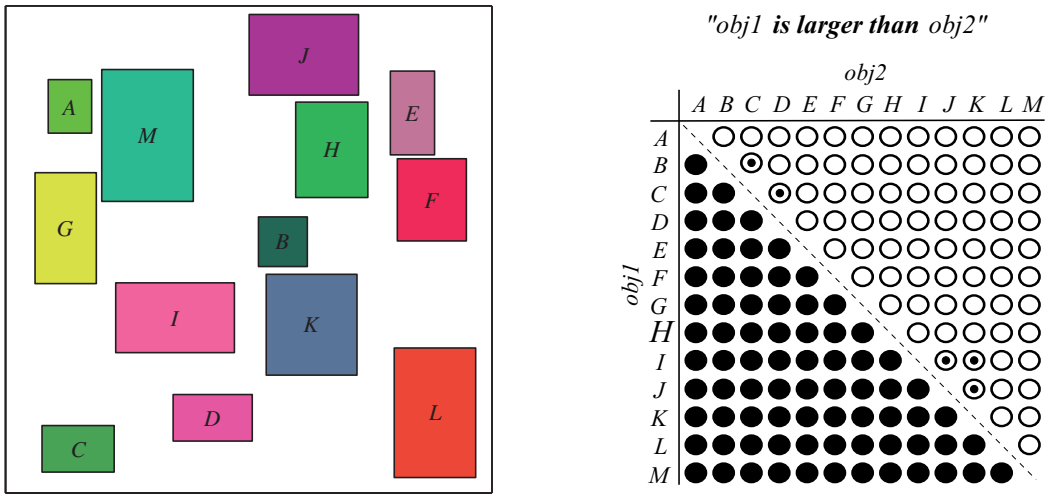


Figure 4.31.: Performance of the system in the word learning scenario: (a) depicts the evolution of the correct categorization rate and (b) that of the NGnet's complexity. For the sake of clarity, the plot only shows results for the learning of three exemplarily selected words.

the initial increase in the system complexity. Since the hidden units serve as prototypes of word-scene associations, most novel observations are correctly categorized in a similarity-based manner. In other words, for any observation the NGnet selects the memorized prototype that best matches the observed scene and finally outputs the corresponding word label. After a while, the hidden units adequately cover upcoming training samples such that those associations do not have to be additionally memorized. The complexity of the NGnet consequently does not further increase; it rather starts to decrease. The reason for this is that enough knowledge has been accumulated such that the feature dimensions, which are most relevant for the word meanings, can be extracted. Thereby, knowledge about the word meaning categories gradually shifts into the extracted features. The complexity of the NGnet consequently decreases and finally remains at a minimal level. Since the extracted features represent the essential aspects of what constitutes a category, a transition from a similarity-based to a rule-based categorization is achieved. This is why the categorization task is also solved more robustly, which reflects itself in a further performance increase towards a near-optimal level.

To provide evidence in favor of a rule-based categorization, the extracted features can be analyzed. More precisely, a rule-based categorization is achieved, if the weight values of the feature extraction matrices $\mathbf{\Phi}$ reflect the real decision criteria. For representing the meaning of terms describing spatial object relations (e.g. *is to the left of* or *is above*) this is the case. The developed word meanings solely relied on the horizontal and vertical relative object positions. For representing relations concerning object sizes (e.g. *is larger than*), the difference in the size of the individual objects has been used. Due to the fact that the feature extraction is linear, an object's size thereby has been approximated by an addition of its width and height (instead of a multiplication). Words related to the brightness of objects (e.g. *is darker than*) are more difficult to represent, since the available RGB color values do not provide the necessary information. Highly non-linear transformations (e.g. to other color spaces) would have to be implemented in this case.



(a) (b)

Figure 4.32.: Application of the acquired word knowledge to the test scene depicted in (a). The model's corresponding categorization of an object being larger than another object is shown in (b). Black circles correspond to category members, white circles to non-members, and dotted circles denote errors made by the system.

Nevertheless, the model approximates this relation sufficiently well by a linear combination of the RGB dimensions, whereas a non-linearity is introduced by an appropriate placement of the NGnet's hidden units. These results show that the built categories solely rely on the core meaning of the corresponding words and therefore demonstrate that the model is able to ground words in the perception of agents.

Finally, the performance of the framework in categorizing an object to be larger than another object is illustrated in Fig. 4.32. Therefore, (a) shows a test scene and (b) the corresponding output of the model when the objects are processed in a pairwise manner. The example demonstrates that the model correctly categorizes most of the observations. Occasional errors only occur for objects with very similar sizes.

4.5. Discussion

In this chapter, a computational model for the grounding of words has been presented. Therefore, word meaning acquisition has been treated as incremental category learning. This is reasonable, since most words refer to scenes of a similar kind or category. The model consequently acquires category representations which correspond to the meaning of words. Thereby, a supervised learning scheme is applied in which words provide labels for the categories to learn. What distinguishes the model from previous approaches is that the proposed framework not only builds categories, but also extracts category-relevant feature dimensions. This feature extraction runs in parallel to category learning and facilitates the categorization task.

The model is developmentally inspired and biologically plausible. Its motivation stems from the fast and slow mapping processes that can be observed in children. Here, it was

argued that CLS theory provides a plausible explanation of these learning patterns and therefore may constitute the biological underpinning of word learning. We consequently endowed the computational model with mechanisms that functionally resemble CLS theory. More precisely, the system comprises a feature extraction layer and a categorization layer which are recurrently coupled. Within the categorization layer an adaptive NGnet serves the incremental category building. Since the NGnet uses localized representations and further possesses mechanisms for an on-demand network growth and pruning, a rapid learning from just few training exemplars is achieved. The NGnet is additionally used to reactivate memorized associations. Samples, which are generated in this way, are used to extract category-relevant features by means of statistical learning techniques.

The framework has been evaluated in various benchmark problems. Thereby, the contributions of the individual processing layers to the overall system performance have been investigated. The results demonstrated that the adaptive NGnet is able to efficiently learn category representations. Its capability of rapid learning is particularly noteworthy. The experiments additionally demonstrated that the incorporation of an information-theoretic feature extraction is beneficial with respect to the generalization capability of the framework. The extracted features cover category-specific information that is used to focus processing on the relevant aspects of inputs. As a consequence, rapid learning is further facilitated, the system complexity is significantly decreased, whereas the system performance stayed approximately the same. A comprehensive comparison to state-of-art approaches showed that the presented model outperforms the other methods in a wide variety of tasks. The application of the model in a word learning scenario finally showed that the framework's learning dynamics are related to those observed in children. Namely, new word-scene associations are rapidly memorized (fast mapping) and acquired knowledge is gradually consolidated (slow mapping). The latter process serves a generalization of initially context-dependent word meanings. This improves the memory representation with respect to both, robustness as well as efficiency.

Finally, problems and restrictions of the framework will be mentioned. These issues partly became evident from the experimental evaluation of the system. The following description includes potential solutions of the problems and hence provides suggestions for future research.

- **Non-Linear Feature Extraction:**
  The feature extraction layer is currently restricted to implement linear transformations, i.e. the feature dimensions are linear combinations of the input dimensions. This limits the kind of information that can be extracted from the inputs. For this reason, a non-linear feature extraction would be desirable. The feature extraction itself does not pose a problem in this respect, since any differentiable function can be learned via stochastic gradient ascent on the mutual information criterion. For example, a *Multi-Layer Perceptron (MLP)* could be trained via backpropagation to extract non-linear discriminative features. The problem, however, arises during the adaption of the NGnet to the changed feature space. Currently, such an adaptation can be done analytically, since any change in the feature space can be described by a linear transformation. For a non-linear feature extraction this would not be the case anymore. For this reason, other adaptation mechanisms need to be found.

- **Memory Reconsolidation:**
  The evaluation of the framework further revealed a significant influence of wrongly given labels on the performance of the system (see Section 4.3.2). This is due to the fact that wrong labels yield an erroneous creation of hidden units which finally disrupt performance. In the current system, such hidden units are only removed if they have turned out to be unnecessary over a long time period. A potential solution to this problem is motivated from the neurobiological process of *reconsolidation*: It is known that any new memory enters a labile state upon reactivation. Reconsolidation describes the process of stabilizing such memories again (Wang and Morris, 2010). The benefit of a labile state is that an erroneously created memory can be rapidly deleted. In detail, a memory item enters a labile state via activation by an input. If the current observation mismatches the memorized association, the item is unreliable. As a consequence, the memory is weakened or even deleted. It has been shown, that the degree of vulnerability depends on the age of a memory, its strength, and the intensity of its reactivation (Alberini, 2011). A similar mechanism could be included in the adaptive NGnet in future. Whenever a hidden unit is activated by an input, the reliability of the memorized association could be checked. Upon mismatch, hidden units could be deleted. Such a process should particularly operate on new memories, whereas old (already consolidated) hidden units should not be affected.

- **Autonomous System Configuration:**
  In Section 4.3.3, the benefits and drawbacks of different system configurations have been discussed. This included the application of global or local NGnets for classification as well as their operation on shared or individual feature spaces. From the discussion it turned out that it is beneficial to consider the learning of one word independent from that of other words, i.e. to use individual classification and feature extraction layers for each word, respectively. As already noted, however, exceptions to this general rule can be found. For example, mutually exclusive words can be represented by the same NGnet, since their category representations do not overlap. Similarly, a shared feature space can be beneficial, e.g. for words that have opposite meanings like *left* and *right*. An automatic construction of suitable system configurations (and possibly a dynamic adaptation of them) is not considered in this thesis. This remains an open issue for future research.

5

Developing Learning Constraints

If names are not correct, language will not be in accordance with the truth of things.

Confucius (551-479 BC)

Learning the meaning of a novel word is a challenging task as the learner initially cannot know to what the word refers to. When hearing a word for the first time, a child is confronted with an indefinite number of potential meanings from which it has to pick the right one. In the previous chapters, different computational models that tackle this problem have been proposed. The model presented in Chapter 3 relies on associative learning theory, insofar as it applies Hebbian learning to map word labels to pre-established potential meanings. Key to the model is the exploitation of the statistical evidence that arises from the observations of multiple distinct word-object pairs. In other words, the model tries to unveil the meaning of a word by discovering what the different word occurrences have in common. The slow learning speed achieved by this model is circumvented by the model proposed in Chapter 4. There, a higher priority is given to word labels insofar as they drive the construction of potential word meanings. This means that in contrast to the associative learning scheme, which used pre-established concepts, this model uses words as supervision signals to construct new concepts (that correspond to the meanings of the words). Inspired by children's fast and slow mapping skills, the model combines different learning techniques and therewith acquires words more rapidly.

The incorporation of learning constraints has not been considered in this thesis so far – or the previous chapters partially took them as granted as we will see later. Literature on word learning, however, assigns a pivotal role to them (cf. Section 2.2). Given the compelling learning speed with which children learn words, it has been argued that

children apply constraints or biases that efficiently restrict the set of potential word meanings and thereby guide language acquisition (Markman, 1990). Examples for such biases are the *whole object assumption* or the *shape bias*. Whereas the *whole object assumption* states that children interpret words as labels for whole objects rather than just parts of them (Mervis, 1987; Hollich et al., 2007), the *shape bias* refers to the fact that the shape of objects, rather than other properties like color, is the primary referent of a word (Imai et al., 1994). If applied in conjunction, both constraints constitute an efficient way of pruning potentially wrong word meaning hypotheses.

It could be shown that even 9-month-old infants exhibit these biases during word learning (Dewar and Xu, 2007). This and other observations let some researchers propose that children are innately equipped with learning constraints (Markman, 1990). As already discussed in Chapter 2, however, there is a lot of controversy about this topic. For example, proponents of the Attentional Learning Account (Smith, 1995) argue that there is no need for innate competencies. Constraints can rather emerge as a byproduct of associative learning and principles of attention (Smith et al., 1996). Furthermore, constraints cannot be fixed over the course of development as they would hinder the acquisition of some words, e.g. those not related to shape. Children consequently have to learn when to trust a bias and when not. At a later age, children are indeed able to consider features other than object shape to be relevant for certain words (Wu et al., 2011). Thereby, caregivers may help to overcome the different learning constraints. For example, Dickinson (1988) showed that the phrase *'made of'* guides children to concentrate on the material of an object rather than its shape.

In this chapter, one of the most studied learning constraints will be discussed – the *mutual exclusivity principle*. The following section will first give a detailed description of the mutual exclusivity principle, highlight its computational importance, and present existing work on this issue. Next, a framework for the development of the learning constraint is presented. Similar to the *Emergentist Coalition Model* (Hollich et al., 2000) we thereby argue for a combination of innately given capabilities, experience-driven adaption, and the incorporation of social-pragmatic cues. We will show how the constraint can efficiently guide word learning by incorporating it into the computational model presented in Chapter 4.

5.1. The Mutual Exclusivity Bias

Mutually exclusive events cannot occur at the same time. With respect to language acquisition the mutual exclusivity principle refers to the fact that children seem to accept only one label per object. This principle becomes evident in studies with children where they are confronted with two objects, a familiar and a novel object. When the children are instructed to *'Give me the x.'*, where x is a nonsense syllable, they will pick the novel object (Markman and Wachtel, 1988). This is reasonable, as they already know a label for the familiar object which differs from x. They consequently think that the novel word has to refer to the unknown object. In this way, the mutual exclusivity principle actively reduces the space of potential word meanings. Together with other learning constraints (e.g. the *whole object assumption* and the *shape bias*) this may allow a child to perform a fast mapping between a word label and its semantic features.

5.1.1. Computational Relevance

Another problem children have to cope with is that caregivers usually only provide positive examples for word meanings. When seeing a cat a mother often says *'Hey, look, there is a cat!'*, whereas she typically does not say *'This is not a dog!'*. Such explicit negative evidence is only provided when parents correct their children (Bohannon and Stanowicz, 1988). A child consequently has to learn the meaning of a word solely based on examples to which the word refers to. For the acquisition of language grammar this issue is also known as *no-negative-evidence problem* (Bowerman, 1988). Learning without negative exemplars obviously is a difficult task as the child has to decide how far a word meaning can be extended. However, the learning constraints children apply are useful in this respect as well. They constitute a source of additional information insofar as negative evidence can be implicitly generated by them. In the following, this is illustrated on the example of the mutual exclusivity principle.

Let us recall the abovementioned scenario where a child is asked to *'Give me the x'*, thereby being confronted with a familiar and a novel object. Since the child selects the novel object to be x, the object certainly constitutes a positive exemplar for the meaning of x. Under the assumption of mutually exclusive labels, however, the familiar object is also of interest. The label x does not refer to the familiar object, such that it can be seen as a negative exemplar for the meaning of x. This implicit generation of negative evidence can be extended to multiple words of course. Any object constitutes a positive exemplar for exactly one word label, but further can be used as a negative exemplar for all other known words. According to Marcus (1993), *"implicit negative evidence thus depends on a reanalysis of positive evidence based on mechanisms internal to the child, rather than input external to the child"*.

However, the mutual exclusivity principle sometimes is misleading and not reliable. An object can obviously have multiple labels, e.g. from different levels of a categorization hierarchy (*animal* and *dog*) or from different languages (*dog* and *hund* in German). A language learner consequently has to decide when to trust the principle and when to break up the assumption. Studies with children (and even adults) revealed that they have problems in accepting multiple labels per object. Only at about the age of 4 years, children's categorization systems seem to be mature enough such that they accept two names for an object, but only if the names refer to different levels of the categorization hierarchy (Au and Glusman, 1990). These findings suggest that the mutual exclusivity principle sometimes hinders early language acquisition. It is therefore unclear whether such constraints are fixed and timely defined or more adaptive and data-driven.

5.1.2. Existing Work

Word learning without explicit negative evidence has been extensively addressed in the work of Regier (Regier, 1990, 1996; Regier and Gahl, 2004). He discussed the no-negative-evidence problem in the context of spatial terms. More precisely, he proposed a computational model which is able to learn spatial relations between a point of interest and a reference square. Thereby, the spatial terms are *above*, *below*, *to the left of*, *to the right of*, *inside*, *outside*, *on*, and *off*.

In contrast to the system presented in Chapter 4, Regier's model comprises a predefined extraction of relevant features as well as a feed-forward neural network to categorize the feature patterns according to the spatial terms. The neural network is trained via backpropagation solely based on positive training exemplars. However, the model includes a mechanism where each training exemplar not only serves as a positive sample for the corresponding category, but also as a negative sample for all other categories to learn. Regier showed that this principle improves performance if reasonable learning rates are chosen. More precisely, the learning rate for negative samples has to be much smaller than that for positive samples. This has two reasons. Firstly, since every training exemplar is a negative sample for all classes except one, negative samples have a much higher impact on learning that needs to be compensated for by the learning rate. Secondly, not all implicitly generated negative training samples are correct. For example, the point of interest can be *outside* the reference square, but at the same time *to the left of* it. In other words, the spatial terms are not all mutually exclusive to each other. The labels rather form word clusters, each of them being composed of mutually exclusive terms (e.g. *to the left of*, *to the right of*, *above*, and *below* as well as *inside* and *outside*).

Using a very small learning rate for the negative samples circumvents the problem of incorrect exemplars, since they only have a small influence on the overall training of the network. However, this only holds as long as the positive exemplars are larger weighted than the negative ones and the number of correct negative samples exceeds the number of incorrect ones. This is a very hard restriction which limits the applicability of Regier's approach. To illustrate this, consider a scenario in which 100 words (partly stemming from different domains) have to be trained. Each positive training sample would implicitly generate 99 negative training samples, many of them being incorrect. For example, a positive example for *large* would be a correct negative sample for *small*, but most probably an incorrect one for *heavy*, *blue*, or *round*. To circumvent this problem, the learning system has to determine which of the words form clusters of mutually exclusive terms and implicitly generate negative evidence only within these clusters (e.g. that an example for *blue* generates negative evidence for *red*, *green*, or *white* but not for *heavy*).

5.2. Our Computational Model

In the visual scene description task, in which we applied our computational model for supervised concept formation (cf. Section 4.4), we considered knowledge on the mutual exclusivity between word labels to be innately given to the system. This means that we designed word classes in which the individual terms either referred to the positions, sizes, or colors of objects. In the simulations, a positive training exemplar for one word finally has been used as negative sample for the other words of the respective class. Children do not possess such an innate knowledge of course. Hence, it is reasonable to model the development of a mutual exclusivity bias during word learning.

It is noteworthy that the approach pursued by Regier (1990), i.e. considering all words to be mutually exclusive, would fail in case of our model. Whereas the arising incorrect training samples are neglectable in the work of Regier, even a few wrongly given labels significantly affect the performance of our model (cf. Section 4.3.2). This is due to the

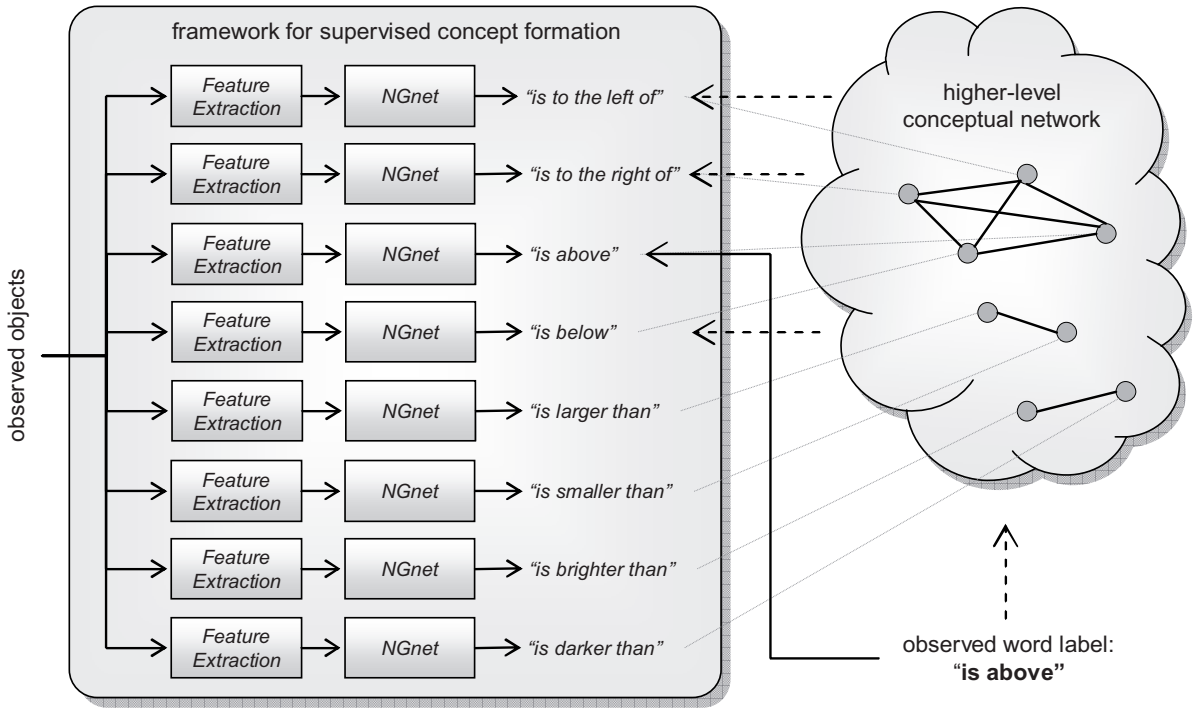


Figure 5.1.: Knowledge on the mutual exclusivity between words may be part of a higher-level conceptual network. Upon hearing the word *is above* the represented word clusters can be used to generate implicit negative evidence for the words of the same cluster, i.e. *is to the left of*, *is to the right of*, and *is below*.

fact, that our model strongly relies on one-shot learning to achieve a rapid word meaning acquisition as compared to Regier's slow learning using large amounts of training data. To make use of implicitly generated negative evidence, the primary task consequently is to determine the words which are mutually exclusive to each other. This is illustrated in Fig. 5.1. If the system is able to identify such word clusters, the problem of wrongly generated negative training samples would vanish.

Here, a computational model for clustering words with respect to mutual exclusivity is proposed. Therefore, it is first discussed which different sources of information a child may use to detect the mutual exclusivity between words. Next, it is proposed how these different cues can be individually estimated and subsequently integrated. Finally, it is shown how word clusters can be built based on the resulting exclusivity measure.

5.2.1. Cues to Exclusive Word Use

In the following it is assumed that a child innately does not have knowledge on which words exclude each other. It consequently has to gain this knowledge over the course of development in order to efficiently use implicit negative evidence. Learning thereby may rely on multiple cues. For example, a child can use parental replies to judge the exclusivity of words (see Marcus (1993) for a review on this topic). Here, we define three cues that in part develop based on the child's increasing word knowledge, in part arise

from the caregiver's attunement to child's utterances, or presuppose cognitive abilities that may be innate to the child. In detail, we define the cues as follows:

- **Overlapping internal word representations:** Word meanings are internally represented in form of categories. These categories entail objects or scenes of a similar kind, each of them being a potential referent of the respective words. If the categories of two words overlap, then there exist objects for which both words constitute an appropriate label. Since such words share referents, they cannot be mutually exclusive. The overlap between the internal word representations consequently provides an intrinsic measure for the exclusivity between words. It is important to note that the reliability of this measure strongly depends on the quality of the built categories. Early in learning the cue will provide uncertain exclusivity estimations, whereas reliability will successively improve over time.
- **Corrections by the caregiver:** The main source of explicitly provided negative evidence are parental corrections of child utterances (Chouinard and Clark, 2003). If a child is corrected, it knows that the previously uttered word was incorrect and, hence, the child can use the observation as a negative training exemplar for the word. What is proposed here is that parental corrections further provide information on word exclusivity. The key idea underlying this proposal is that caregivers typically correct their children by using labels that stem from the same domain. For example, if a child wrongly names a person, a mother provides the correct name of the person. If the child wrongly labels an object color, the mother says the correct color. But she typically does not correct the child by using a word related to object shape. The two words that have been uttered by the child and the mother consequently are mutually exclusive.
- **Semantically rich scene descriptions by the caregiver:** A caregiver often not only uses one word to refer to an object (e.g. *ball*), but embeds the label into phrases (e.g. *the big blue ball*) that allow the child to determine the referent more easily. In these semantically rich utterances multiple words are used to describe an object. Since these words share the referent, they cannot be mutually exclusive. Multi-word descriptions hence provide evidence against an exclusive word use.

5.2.2. Cue Estimation

Each of the abovementioned cues provides information on the pair-wise exclusivity of words. This means that the cues do not yield clusters of exclusive words (e.g. color terms), such clusters rather can be subsequently constructed on the basis of the pair-wise estimates (e.g. *red-blue*, *red-green*, *blue-yellow* etc.) Here, we first show how pair-wise exclusivity measurements can be obtained.

The latter two cues, i.e. evidence arising from corrections or multi-word phrases, can be easily evaluated. Both cues yield word-pairs that are either exclusive or non-exclusive. These observations can be memorized in a look-up table, e.g. a matrix, whose cells correspond to all possible pair-wise combinations of words. If the caregiver corrects the word *blue* with the word *red*, for example, the system knows that *blue* and *red* are mutually exclusive and can memorize this measurement by assigning 1 to the corresponding

c_i \ c_j	+1	-1
+1	*a*	*b*
-1	*c*	*d*

Figure 5.2.: A 2x2 contingency table.

matrix cell. In contrary, a cell is filled with 0 in case of a non-exclusive word-pair. Of course, evidence could be accumulated over multiple training exemplars, thereby yielding continuously valued matrix entries. However, here we stick to the binary values.

The estimation of the first cue is less obvious. This is because the internally built word categories can possess complex shapes such that an overlap between them cannot be calculated analytically. For this reason, we pursue a sample-based strategy to estimate the category overlaps. Given a large amount of randomly sampled input patterns, the built categories are used to estimate whether the patterns are members or non-members of the categories. More precisely, for each input pattern $\boldsymbol{x}$ the internal representation of a word w_i returns a binary decision $c_i(\boldsymbol{x}) \in \{-1, +1\}$ indicating whether the word w_i is appropriate for the description of $\boldsymbol{x}$. The overlap between two word categories finally can be estimated by comparing the respective binary responses. Largely overlapping (and therewith non-exclusive) categories should result in multiple co-occurrences of $+1$, since the words share similar referents. In contrast, word exclusivity is characterized by a response pattern, where the respective categories do not return $+1$ simultaneously.

The most common techniques for measuring binary response pattern similarity rely on a 2x2 contingency table. Thereby, a contingency table is a matrix whose entries refer to the frequency of different response combinations of two categorical variables. In the example depicted in Fig. 5.2, a denotes the proportion of patterns for which both word categories output $+1$, b and c refer to the proportion of patterns for which just one category outputs $+1$, and d is related to those patterns that evoke a response of -1 for both words. Accordingly, we fill the table as follows.

$$\begin{aligned} a &= \sum_{\boldsymbol{x}} p(c_i = +1|\boldsymbol{x}) \cdot p(c_j = +1|\boldsymbol{x}) \\ b &= \sum_{\boldsymbol{x}} p(c_i = +1|\boldsymbol{x}) \cdot p(c_j = -1|\boldsymbol{x}) \\ c &= \sum_{\boldsymbol{x}} p(c_i = -1|\boldsymbol{x}) \cdot p(c_j = +1|\boldsymbol{x}) \\ d &= \sum_{\boldsymbol{x}} p(c_i = -1|\boldsymbol{x}) \cdot p(c_j = -1|\boldsymbol{x}) \end{aligned} \tag{5.1}$$

Thereby, the individual probabilities stem from the outputs of the NGnets which represent the respective word meaning categories.

The entries of the contingency table can be used to estimate response similarity. Thereby, existing similarity indices can be grouped into those measures that incorporate the value d

and those that do not consider d (see the work of Warrens (2008) for a review of different indices). For estimating the mutual exclusivity between two words, it is important not to incorporate d. This is due to the fact that d refers to situations in which an object is observed, but both words do not provide an appropriate label for the object, e.g. *'cat'* and *'dog'* do not serves as a label for a *tree*. Such situations do not provide evidence with respect to word exclusivity and hence should not be considered. Here, we estimated the exclusivity $excl_{ij}$ between two words w_i and w_j according to

$$excl_{i,j} = \left[\frac{b \cdot c}{(a+b) \cdot (a+c)} \right]^{\frac{1}{2}}, \tag{5.2}$$

which is related to the similarity index proposed by Ochiai (1957). As can be seen from the formula, $excl_{ij}$ is close to 1 if a is small in comparison to b and c. In contrast, if two words share many referents, i.e. a is large, $excl_{ij}$ will be close to 0. We further set the diagonal elements of matrix $Excl = \{excl_{ij}\}$, i.e. the exclusivity between a word and itself, to 1. Even though a word is not exclusive to itself, this is reasonable as we will show later.

5.2.3. Cue Integration

As it is illustrated in Fig. 5.3, the three different cues yield exclusivity estimates for word pairs. We memorize these estimates in matrices whose entries refer to all different combinations of words. Our aim is to construct word clusters in a way that words within a cluster are mutually exclusive whereas words of different clusters are not. To do so, the measurements arising from the different cues are first integrated. In the resulting matrix, the sought word clusters should pop out insofar as exclusive words exhibit correlated pair-wise exclusivities to other words. A pair-wise exclusivity matrix of this kind can consequently serve as the basis for word clustering. The clustering of words will be in the focus of the next section. Here, we first present a model for cue integration.

The integration of the different cues constitutes a difficult task for two reasons. Firstly, the individual measurements are noisy. In particular, estimates arising from the overlap of word categories can be noisy, since this cue presupposes that the categories already have been correctly built. This is obviously not the case at the beginning of the training. Secondly, cue integration suffers from missing values. Especially the matrices estimated based on tutor corrections or multi-word phrases are affected in this respect. This is because the matrix entries are filled according to the words that have been uttered. If certain word pairs did not appear in the conversation, the corresponding matrix cells remain empty. To cope with these constraints, we chose to use a *Markov Random Field (MRF)* (Kindermann and Snell, 1980). Until now, MRFs have been mainly applied in computer vision tasks (Li, 1995), e.g. for image segmentation. Due to their computationally attractive features, however, here we adopt MRFs for the purpose of cue integration. In detail, MRFs possess a probabilistic computation and hence are suited for handling noisy data. Furthermore, MRFs can be used to estimate hidden variables via Bayesian inference and consequently can cope with missing values.

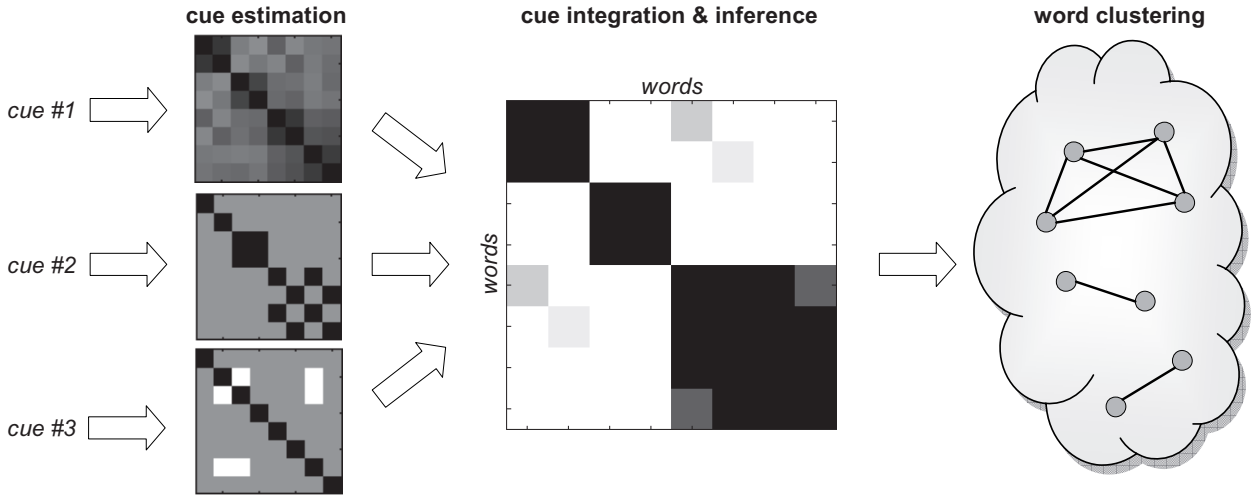


Figure 5.3.: Pair-wise word exclusivities are first estimated individually based on multiple cues. The results are subsequently integrated and finally used for word clustering. Each element of a matrix refers to the exclusivity of a particular word pair. Dark colors indicate high exclusivity, whereas light colors denote non-exclusivity.

A MRF can be described by an undirected graph $\mathcal{G} = (V, E)$ composed of vertices V and edges E. Thereby, the vertices denote the variables to be estimated, whereas the edges express dependencies between them. During computation, a potential is assigned to each vertex. These potentials can spread within the field via the edges. In our model, the node set is given by $V = \{v_{ij}\}$ with $i, j = 1, \ldots, N$ where N is the number of words. The vertex v_{ij} consequently refers to the word pair composed of the words w_i and w_j. The potential $pot(v_{ij})$ is the integrated word exclusivity $\widehat{excl}_{ij}$ we finally would like to calculate. We initialize the node potentials according to

$$pot(v_{ij}) = \frac{1}{M_{ij}} \cdot \sum_{k=1}^{3} excl_{ij}^{(k)}, \tag{5.3}$$

where $excl_{ij}^{(k)}$ denotes the exclusivity estimate from the k-th cue, $M = 3$ is the number of cues, and M_{ij} is the number of valid cues. This means that cues with missing entries for element ij where excluded in Eq. (5.3). In case of all cues being invalid, i.e. $M_{ij} = 0$, we set $pot(v_{ij}) = 0.5$.

If we consider the vertices V to be arranged in a way that they resemble the elements of the target matrix $\widehat{Excl}$, then we include edges that allow the node potentials to spread horizontally and vertically (see Fig. 5.4). Thereby, an undirected edge between two nodes v_{ij} and v_{kl} is denoted by e_{ij-kl}. It is important to note that self connections are excluded, i.e. $v_{ij} \neq v_{kl}$. Each edge features a weight that has been chosen based on the initial potentials of the two vertices that are connected by the edge. In detail, we set

$$e_{ij-kl} = [pot(v_{ij}), 1 - pot(v_{ij})]^T * [pot(v_{kl}), 1 - pot(v_{kl})]\,. \tag{5.4}$$

As a result, an edge features a large weight, if it connects two mutually exclusive word pairs or two non-exclusive word pairs. In contrast, small weights are assigned to edges between an exclusive and a non-exclusive word pair.

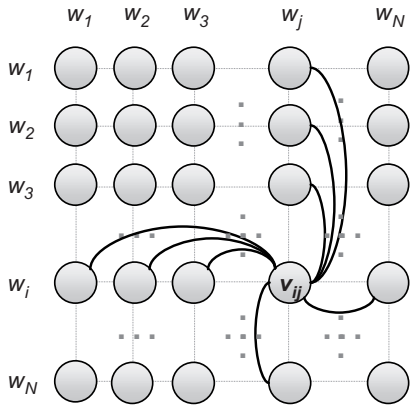


Figure 5.4.: The connectivity of the Markov Random Field.

We finally applied *Loopy Belief Propagation* to spread the potentials within the network. Thereby, mutually exclusive word pairs will reinforce each other due to their large connection weights. The same is true for non-exclusive word pairs. However, the primary aim of Loopy Belief Propagation is to infer the exclusivity of word pairs for which the different cues resulted in missing or uncertain values. To do so, the chosen node connectivity implements a transitive rule of the form: if the words w_i and w_j are mutually exclusive and the words w_j and w_k are mutually exclusive, then w_i and w_k are mutually exclusive, too. This is due to the fact that both vertices v_{ij} and v_{jk} propagate their large potentials to node v_{ik}. Fig. 5.3 depicts a concrete example in which the evidences arising from the three different cues have been integrated using the aforementioned technique. As can be seen, the noisy pattern resulting from the first cue gets much more prominent by incorporating the partial biases provided by the second and third cue.

5.2.4. Word Clustering

Clusters of mutually exclusive words distinguish themselves as relatively homogeneous subgroups in the inferred exclusivity matrix $\widehat{Excl}$ (as it is the case in Fig. 5.3). However, it is important to note that these subgroups are not always characterized by homogeneous regions within the matrix, since this requires an appropriate ordering of the matrix rows and columns. Additionally, the estimated exclusivities still can be very noisy, particularly at the beginning of training. These aspects render the clustering of words a rather challenging task.

We solve this task by introducing a *mutual exclusivity space* in which the N different words can be placed. The space comprises N dimensions, each of them denoting the pair-wise exclusivity to one particular word. This means that we can interpret the columns of the matrix $\widehat{Excl}$ as the position vectors of the N words with respect to the *mutual exclusivity space*. As a consequence, words that show correlated exclusivities are positioned close-by, while they feature large distances to the other words. This finally allows us to use standard distance-based techniques for clustering the words. It is noteworthy that by relying on this interpretation of data, we further bridge the gap between pair-wise exclusivity estimates (as obtained by integrating the different cues) and a measure for the exclusivity between

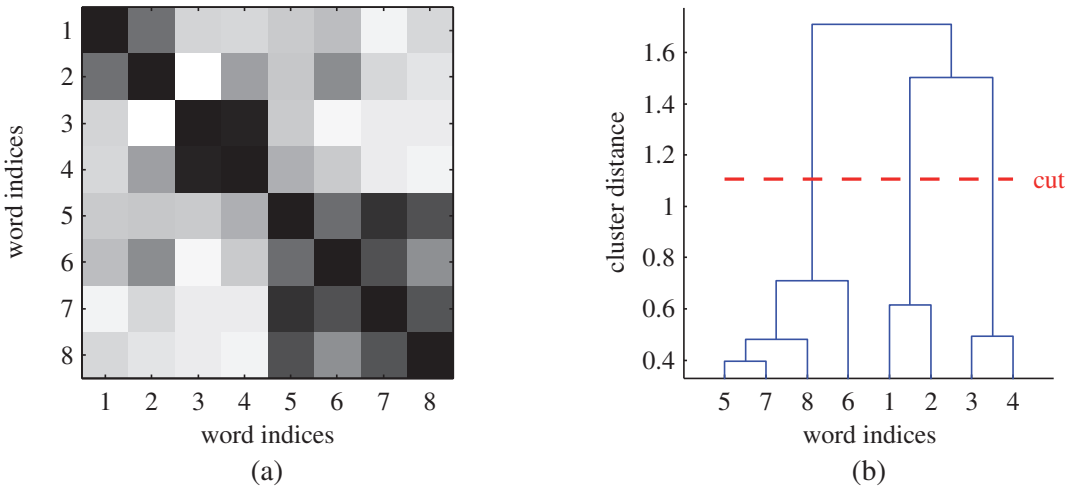


Figure 5.5.: The hierarchical clustering of the input data shown in (a) yields the word clusters depicted by the dendrogram in (b).

multiple words. This is because two words have to possess similar pair-wise exclusivity values with respect to all other words in order to be positioned close-by.

We use *agglomerative hierarchical clustering* (Duda et al., 2000) to group the different words. This technique initially assigns each point in space to a separate cluster, i.e. N words result in N clusters. Subsequently the number of clusters is reduced by merging the two most similar clusters. This step is finally repeated until all words belong to the same cluster. For merging the clusters different distance metrics and linkage methods are available (Duda et al., 2000; Quackenbush, 2001). Here, merging is carried out based on the *Euclidean distance* and *average linkage*. In detail, this means that the distance between two clusters is calculated as the average of the Euclidean distances between each point in the first cluster with all other points in the second cluster. In Fig. 5.5 the clustering solution for a particular example is shown. Thereby, (a) depicts an exclusivity matrix used as input, whereas (b) displays the formed clusters in a dendrogram.

What remains is to choose the number of clusters. As illustrated in Fig. 5.5 (b), this corresponds to finding a cluster-distance level at which the dendrogram should be horizontally cut. A word cluster then comprises all words that are part of the same branch below the cut. Many methods exist for calculating an optimum number of clusters k (see Milligan and Cooper (1985) for a review). They all rely on error criteria that express how good the clustering solutions for different values of k are. Thereby, the within-cluster sum of squared distances is most commonly used. It is defined as

$$W(k) = \sum_{l=1}^{k} \left[\frac{1}{2n_l} \cdot \sum_{i,j \in C_l} d_{i,j}^2 \right], \tag{5.5}$$

where $C_1, \ldots, C_k$ denote the k calculated clusters that each comprise the indices of the respective word samples, n_l is the number of words in cluster C_l, and $d_{i,j}$ refers to the Euclidean distance between the words with indices i and j. We hence obtain $d_{i,j}$ by comparing the i-th and j-th column of the exclusivity matrix $\widehat{Excl}$.

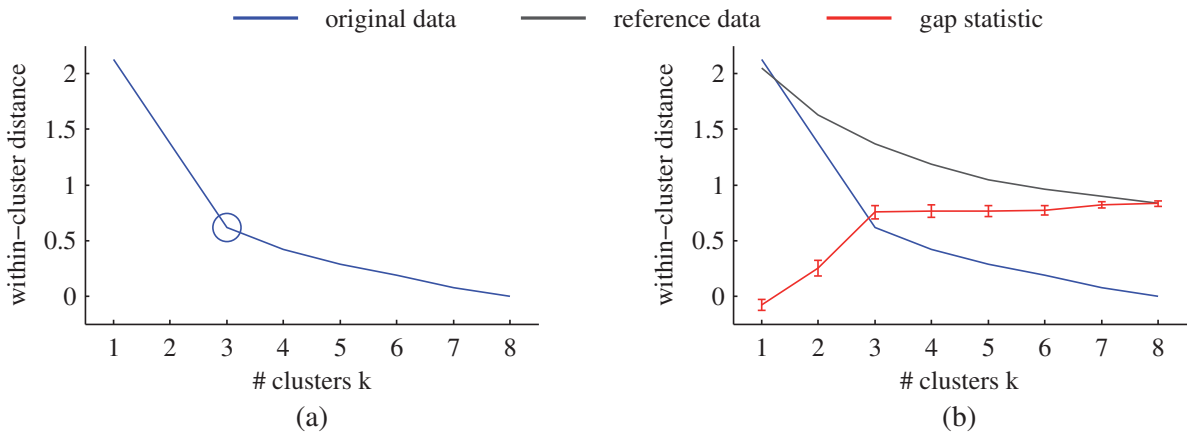


Figure 5.6.: In (a) the within-cluster sum of squares is plotted as a function of the number of clusters k. The circle marks the optimum $\widehat{k} = 3$. The gap statistic is calculated by comparing the function in (a) with one obtained from clustering uniformly distributed reference data. The resulting criterion is depicted in (b).

In Fig. 5.6 (a) the within-cluster sum of squares is plotted for the example depicted in Fig. 5.5. As can be seen, the measure monotonically decreases as the number of clusters increases. This is due to the fact that a more fine-grained clustering can be achieved when a larger k is used. However, the criterion rapidly decreases only up to a certain value $\widehat{k}$ (marked by the circle). For values $k > \widehat{k}$ the measure $W(k)$ just slightly decreases. Existing methods aim at detecting such an "elbow" in the function W as it marks the number of clusters that partition the dataset optimally. In reality, however, the "elbow" most often is less prominent than depicted in the example. It consequently is difficult to detect by the different methods.

The *gap statistic* proposed by Tibshirani et al. (2001) tries to overcome this issue by comparing the obtained $W(k)$ to the within-cluster sum of squares $W_{ref}(k)$ resulting from the clustering of a reference distribution. More precisely, reference data points are uniformly sampled from a rectangular region that constitutes the bounding box of the original word samples. Thereby, the bounding box is oriented along the principle dimensions of the data. The underlying idea is that the grouping of clustered data (the original words) should result in smaller within-cluster sum of squares as compared to partitioning uniformly distributed data. The gap statistic consequently calculates the gap between the two criteria according to

$$gap(k) = W(k) - W_{ref}(k) \tag{5.6}$$

and determines the optimal number of clusters $\widehat{k}$ as the minimum k that satisfies

$$gap(k) \geq [gap(k+1) - s_{k+1}]\,. \tag{5.7}$$

Here, s_{k+1} is a value that is proportional to the standard deviation of $gap(k+1)$ as obtained from clustering multiple reference sample sets (see Tibshirani et al. (2001) for a detailed description). Fig. 5.6 (b) illustrates that by using the gap statistic the optimum number of clusters ($\widehat{k} = 3$) can be reliably estimated for the example data.

5.3. Evaluation in a Word Learning Scenario

To assess the performance of the proposed computational model, it has been incorporated into the word learning framework presented in Chapter 4. For the evaluation we further use the same visual scene description task (see Section 4.4) in order to obtain comparable results. More precisely, the system has to learn the meaning of words for the relations between geometric objects. This includes words for relations concerning object positions (*is to the left of*, *is to the right of*, *is above*, *is below*), object sizes (*is larger than*, *is smaller than*), or object colors (*is brighter than*, *is darker than*). But in contrast to the previous simulations, knowledge on word exclusivity is not innately given to the system. It rather has to learn that the words stem from different domains (position, size, and color), insofar as some words can simultaneously be used for the description of a scene whereas others cannot. The proposed model should learn clusters which comprise words that are mutually exclusive to each other. Words of different clusters should be non-exclusive and hence can serve as labels for the same observation. Consequently, in the simulations we implicitly generate negative training exemplars for a word meaning only from samples that constitute positive training exemplars for words of the same cluster. In summary, the experiment differs from the one presented in Section 4.4, insofar as clusters of exclusive words are learned rather than being predefined.

During learning the tutor selects two objects and describes the relation between them, e.g. by saying *'This object is above that object'*. He hence presents a positive training exemplar for the word meaning. In Chapter 4, the simulations were solely based on this kind of interaction. The presentation of correct word labels constitutes the standard interaction paradigm here as well. However, we further increase the diversity of conversation patterns by including other interaction modes. These modes allow the system to base learning on each of the previously mentioned cues for word exclusivity. In detail, the following interaction modes are additionally applied:

(1) **Wrong word label:** The tutor uses a wrong label to highlight that a word cannot be used for the description of a certain scene, e.g. by saying *'This object is not to the left of that object'*. The corresponding observation hence can be used as a negative training exemplar for the word meaning. Even though this type of conversation seldomly appears in mother-child interaction, we included it in the present evaluation.

(2) **Correction:** The system is asked to describe a scene but erroneously uses an incorrect label. The tutor subsequently corrects the system, e.g. by saying *'No, this object is not to the left of that object. It is above the object'*. The learner hence is provided with a negative training exemplar for the meaning of the wrongly uttered word as well as a positive training exemplar for the meaning of the correct word. Moreover, the system can use this conversation to gain knowledge concerning the exclusivity of the two words.

(3) **Multi-word description:** The tutor describes the relation between objects by using multiple words, e.g. by saying *'This object is above and larger than that object'*. The observation hence serves as a positive training exemplar for the meaning of multiple words. In addition, the learner subsequently knows that the words can occur simultaneously and hence are non-exclusive.

It is noteworthy that, as in Chapter 4, here we consider the learner to possess sufficient knowledge for segmenting the utterances. This means that the learner is able to determine the words of interest, such that the respective network parts can be trained. In the following, the computational model is thoroughly evaluated. Thereby, a special emphasis is first given on how well the model is able to acquire the clusters of mutually exclusive words and, second, what the individual contributions of the different interaction modes are. We finally compare the obtained results to those of the previous chapter.

5.3.1. Word Clustering Performance

To evaluate the clustering of mutually exclusive words, we run a simulation in which different conversation patterns have been sequentially presented to the system. These samples were randomly chosen from a training data set. Thereby, it has been taken care that each interaction mode is used equally often on average. This means that the system is provided with a single positive training exemplar, a single negative training exemplar, a correction pattern, or a multi-word description in 25% of the samples, respectively.

At each instance in time, the developed word clusters were recorded. The quality of a clustering solution has been calculated on the basis of exclusivity estimates for all possible word pairs. In detail, we checked whether two mutually exclusive words have been assigned to the same cluster or, alternatively, whether two non-exclusive words have been assigned to different clusters. The ratio between the number of correctly assigned word pairs and the total number of word pairs finally yields a quality index ranging from 0 to 1. Thereby, a value of 1 denotes a totally correct cluster solution. Fig. 5.7 depicts the resulting cluster qualities. As can be seen, the system is able to correctly cluster all words after the presentation of approximately 100 training samples. However, mutual exclusivity has been correctly estimated for a large proportion of word pairs ($\approx 80\%$) already after just 20 training samples. This demonstrates that the system is able to rapidly develop clusters of exclusive words. A rapid learning is particularly important, since the word clusters are used to implicitely generate negative training samples for word learning. Incorrect cluster solutions can result in wrongly generated training samples which may degrade system performance.

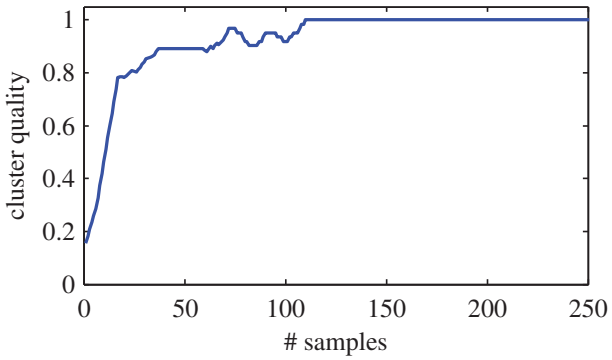


Figure 5.7.: The quality of the developed word clusters as a function of the number of training samples that have been presented to the system.

5.3.2. Individual Contributions of the Interaction Modes

Given the compelling speed with which correct word clusters are acquired, the question arises whether certain interaction modes are particularly important for learning. More precisely, we would like to answer the following questions: What kind of tutor-system interaction is necessary to efficiently learn? Are some interaction modes more important than others? Or are the different conversation patterns equally relevant? Answering these questions is of interest for two reasons. Firstly, it suggests how a user should teach the system in order to aid learning. Secondly, it not only unveils relevant interaction modes, but also the cues that most efficiently guide word clustering. This is due to the fact that the different cues rely on different interaction patterns (cf. Section 5.2.1).

We run multiple simulations to estimate the individual contributions of the different interaction modes. Thereby, the simulations differed in the amount of training samples that have been generated using the different modes. In detail, let p_1, p_2, and p_3 denote the proportion of samples stemming from the different modes, respectively. We further assume that we want to estimate the contribution of interaction mode #1. Then we first set $p_2 = p_3 = 25\%$ and vary $p_1 = 0 \ldots 25\%$. By doing so, the relevance of mode #1, given the presence of the other modes #2 and #3, can be estimated. Similarly, we can estimate the relevance of mode #1, given the absence of modes #2 and #3, by setting $p_2 = p_3 = 0\%$ and varying p_1. For each parameter setting, the remaining amount of training samples $(100\% - (p_1 + p_2 + p_3))$ is generated using the standard interaction paradigm, i.e. by presenting a single positive training exemplar.

In the following, we first focus on the contribution of interaction mode #1, i.e. the explicit presentation of negative training exemplars. Fig. 5.8 therefor depicts the evolution of cluster qualities for simulation runs in which different values of p_1 have been chosen. In (a) we set $p_2 = p_3 = 25\%$, whereas (b) shows simulation results using $p_2 = p_3 = 0\%$. As can be seen, the plots in (a) do not differ much. This suggests that the first interaction mode

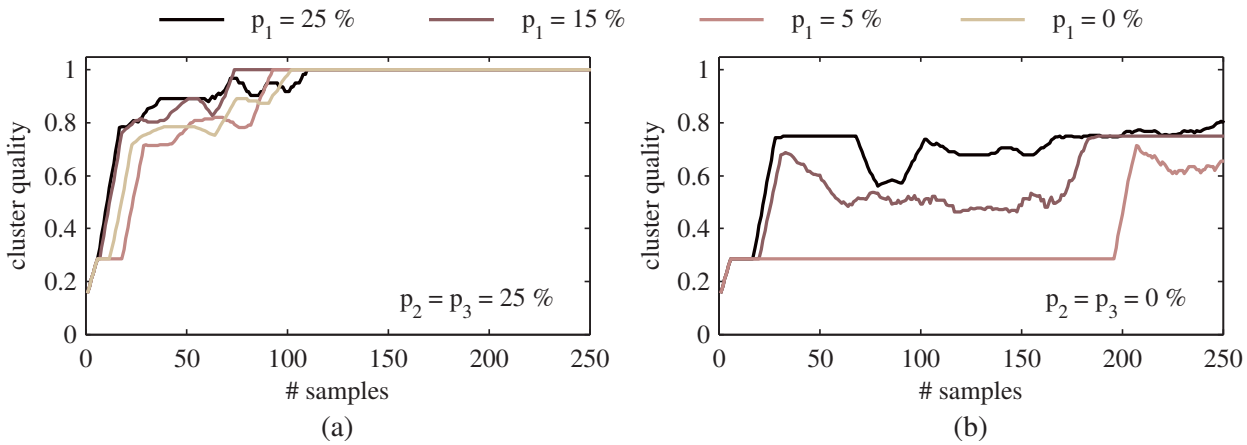


Figure 5.8.: The evolution of cluster qualities in simulations with varying amounts p_1 of explicitly presented negative training exemplars. In (a) results are depicted for simulations in which the other interaction modes were present ($p_2 = p_3 = 25\%$), whereas they were absent ($p_2 = p_3 = 0\%$) in the simulations depicted in (b).

does not have a significant influence on word clustering, if training samples are generated by the other interaction modes, too. The minor influence of mode #1 is supported by the results depicted in (b). There it can be seen that explicitly generated negative training samples alone (i.e. in the absence of the other interaction modes) are not sufficient for developing appropriate word clusters.

The contrary is true concerning interaction mode #2, i.e. conversations in which the tutor corrects a wrong labeling by the system. The corresponding results are depicted in Fig. 5.9. From the plots in (a) it becomes evident that correction patterns are important for developing appropriate word clusters even in the presence of the other interaction modes. The system's capability to detect word exclusivity decreases if the amount of corrections by the tutor decreases. Furthermore, (b) illustrates that learning solely based on interaction mode #2, i.e. in the absence of the other modes, is successful if a sufficient amount of corrections is provided by the tutor.

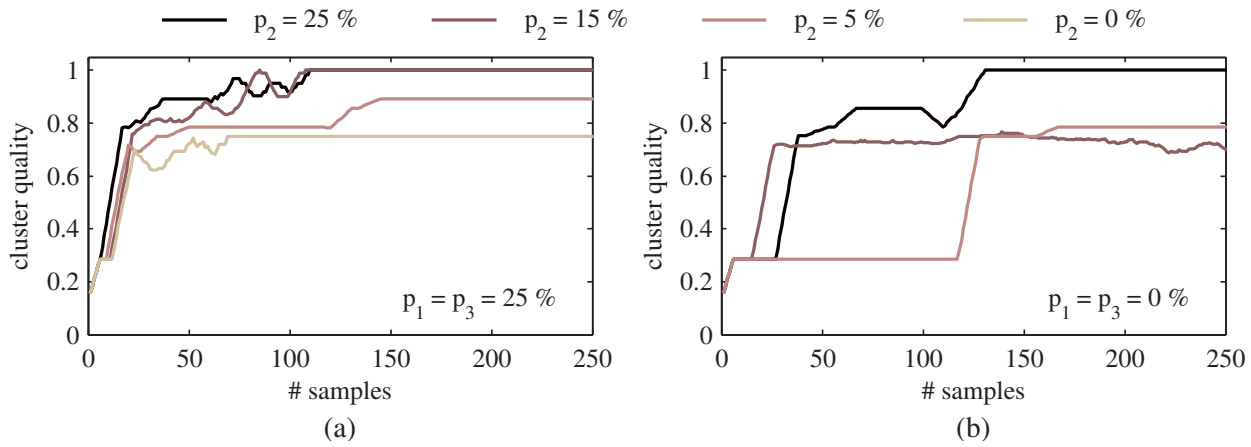


Figure 5.9.: The evolution of cluster qualities in simulations with varying amounts p_2 of corrections by the tutor. In (a) results are depicted for simulations in which the other interaction modes were present ($p_1 = p_3 = 25\%$), whereas they were absent ($p_1 = p_3 = 0\%$) in the simulations depicted in (b).

The relevance of multi-word descriptions can only be evaluated in the presence of the other two modes. This is due to the fact that multi-word descriptions only yield positive training exemplars and further do not provide evidence on which words are exclusive to each other. As a consequence, the word learning framework would not obtain negative training exemplars – neither explicitly provided by a tutor nor implicitly generated based on mutual word exclusivity. The cluster qualities obtained when the other two modes are present ($p_1 = p_2 = 25\%$) are shown in Fig. 5.10. The plots demonstrate that the third interaction mode is not necessary for the development of correct word clusters. However, it seems to be advantageous, insofar as a faster learning is achieved when more multi-word description are supplied to the system.

Overall, the evaluation revealed different individual contributions of the interaction modes: Firstly, the explicit presentation of negative training exemplars (e.g. *This object is not to the left of that object.*) is not necessary for word learning. This is in accordance

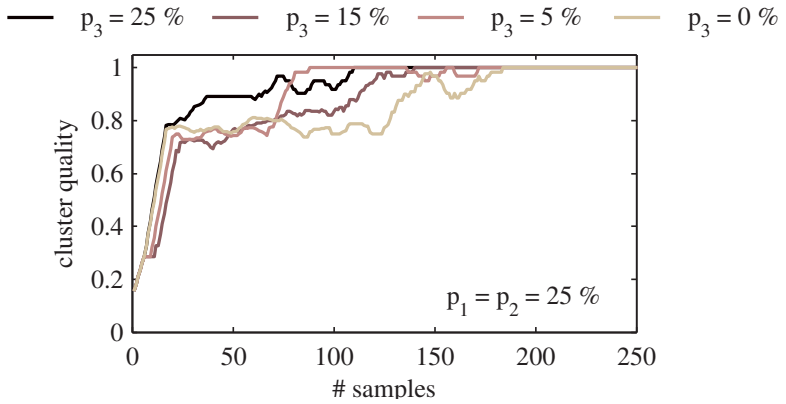


Figure 5.10.: The evolution of cluster qualities in simulations with varying amounts p_3 of multi-word descriptions. Results are depicted for simulations in which the other interaction modes were present ($p_1 = p_2 = 25\%$).

with the fact that caregivers seldomly supply this kind of information to their children. Secondly, corrections of erroneously uttered words are of particular importance for word learning. Evidence in favor of this is provided by studies on mother-child interaction. Chouinard and Clark (2003) showed that mothers most often correct wrong utterances of their children. The children in turn seem to strongly use this kind of information during word learning. Finally, multi-word description are not necessary, but beneficial, for word learning. This is supported by a study of Weizman and Snow (2001) who showed that the embedding of words into semantically rich descriptions facilitate the acquisition of the words by children.

5.3.3. Overall System Performance

The previous simulations assessed the system's capability to learn clusters of mutually exclusive words. What remains is to evaluate the performance with respect to word meaning acquisition. Both aspects are certainly intertwined. This is due to the fact that the word clusters are used to implicitly generate negative training exemplars which are subsequently used for word learning. Incorrect word clusters can consequently yield wrong training samples which finally may disrupt system performance. However, it is important to note that incorrect word clusters not always yield incorrect training samples. For example, assume an incorrectly built cluster that is composed of just one word, i.e. the word has been estimated to be non-exclusive to all other words. No negative evidence will be generated in this case, since mutual word exclusivity is a prerequisite to do so.

To investigate the effect of word clustering, we compared the results of simulations in which either learned or predefined word clusters were used to implicitly generate negative training exemplars. Similar to Section 4.4, the evolution of the categorization error and the evolution of the network size were recorded for each simulation. By comparison of the different result curves, the effect of word clustering on system performance, system complexity, and learning speed can be estimated. Overall, the results showed qualitatively

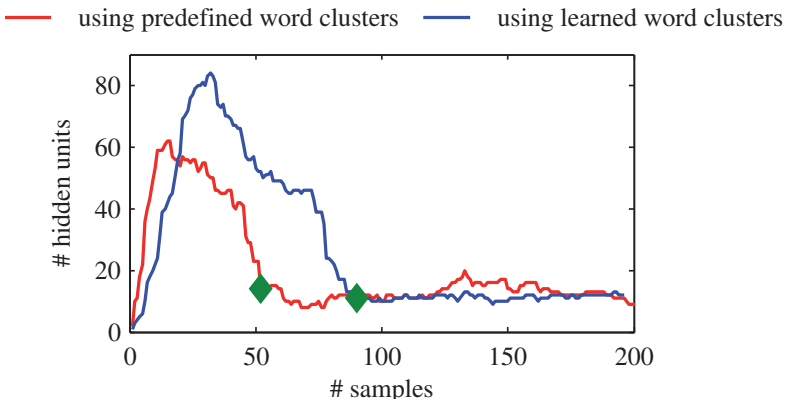


Figure 5.11.: The plots illustrate how network size evolved during the learning of *is below*. In one simulation, clusters of mutually exclusive words have been learned, whereas the other simulation relied on predefined word clusters. Green diamonds mark instances in time at which network training reached convergence.

similar learning patterns. This is exemplarily depicted in Fig. 5.11 where the evolution of the system complexity is shown for the learning of *is below*. As can bee seen, the plots exhibit similar characteristics insofar as network size initially increases, subsequently decreases, and finally converges to a minimum level. Even though the curves are qualitatively similar, it is difficult to assess this similarity quantitatively. This is due to the fact that a temporal offset between the curves exist. For this reason, an alignment has been carried out, insofar as the time instances at which the results reached convergence were extracted for all curves, respectively. In Fig. 5.11 these time instances are marked by a green diamond. Finally, we calculated the following quantitative measures:

- **Difference in learning speed:** The temporal offset between two convergence points is taken as an indicator on how much word clustering affects learning speed as compared to using predefined word clusters.
- **Difference in system performance:** The average difference between the achieved categorization rates is considered an indicator for the effect of word clustering on system performance. Thereby, we only consider the categorization rates that have been achieved after the training of the system converged, i.e. data points after the extracted convergence points. This is reasonable as we are interested in the final performance of the system.
- **Difference in system complexity:** The average difference between the network sizes is used as a measure on how much word clustering affects system complexity as compared to relying on predefined knowledge on word exclusivity. For the same reason as above, only data points after convergence are taken into account.

The abovementioned measures have been calculated for multiple simulations in which different interaction modes were applied for constructing training samples. The results obtained from a comparison of the individual simulations with a run, in which word clusters were predefined, are summarized in Table 5.11. There, the results are averaged over the different words that have been learned.

applied interaction modes	difference in learning speed [# samples]	difference in system performance [%]	difference in system complexity [# units]
1	22.0	-0.62	-0.37
2	7.6	0.44	-0.55
3	—	—	—
1 + 2	1.8	0.15	-0.35
1 + 3	41.4	-0.38	-0.06
2 + 3	13.6	0.27	-0.01
1 + 2 + 3	8.4	0.33	-0.08

Table 5.1.: The differences in the results obtained when clusters of exclusive words are learned versus predefined. Word clustering has been assessed in multiple simulations using different interaction modes.

As can be seen from the results, the learning of word clusters neither affects system performance nor system complexity. Irrespective of the applied interaction modes, the sizes of the trained networks differed in less than one hidden unit. Similarly, each simulation achieved categorization rates which differed less than 1% from the results obtained when using predefined word clusters. Both, the effect on system performance as well as the effect on system complexity, hence can be neglected. However, the development of word clusters affected learning speed, insofar as more training samples were needed to reach convergence. In this respect, the results further suggest that learning speed is influenced by the applied interaction modes. Particularly the incorporation of the first and the third mode seem to slow down learning. In contrast, corrections supplied by a tutor (mode #2), did not affect learning speed so much. The results hence provide further evidence for a special importance of corrections during word learning. Tutor corrections not only efficiently guide word clustering (cf. Section 5.3.2), but also yield learning speeds that are close to those achieved when using predefined word clusters.

5.4. Discussion

After having presented computational models for word meaning acquisition using both unsupervised and supervised learning techniques, the aim of this chapter was to investigate constraints that may guide word learning. In previous work, many different learning constraints have been suggested (Markman, 1990). Whereas there is no doubt that children's word learning relies on such biases, there is controversy on whether the constraints are innate or develop through learning (Markman, 1994; Smith et al., 1996). Our investigation focused on the mutual exclusivity bias, which constitutes one of the constrains that have been most extensively studied so far. We illustrated that this bias is of particular computational importance, as it leads a way to overcome the *no-negative-evidence problem*. More precisely, mutual word exclusivity can be used to implicitly generate negative training exemplars during word learning. Such negative evidence on

the contrary is seldomly supplied by caregivers. We also pointed out the problems that an innate existence of the bias would come with.

To overcome these issues, we suggested that the bias can develop on the basis of minimum innate knowledge as well as social-pragmatic cues available during mother-child interaction. More precisely, a computational model has been presented that is able to form clusters of mutually exclusive words. These clusters can be used to implicitly generate negative training exemplars, thereby guiding word learning. In detail, the model first estimates pair-wise word exclusivities on the basis of three cues. These cues rely on knowledge that either is internal to the system (e.g. overlapping word representations) or explicitly provided by caregivers during interaction (e.g. corrections or multi-word description). The evidences arising from the individual cues are subsequently integrated and finally used to hierarchically cluster the words.

Our experimental evaluation showed that a mutual exclusivity bias indeed does not have to be innate to a system, but rather can develop over the course of learning. The effect of learning as compared to using predefined knowledge is neglectable, insofar as the system acquired word meanings equally well. However, it could be shown that an appropriate development of the bias strongly depends on the kind of tutor-system-interaction that is used during word learning. More precisely, utterances in which a tutor corrects system mistakes seem to be of particular importance. On the contrary, the explicit presentation of negative training exemplars as well as the embedding of words into semantically rich descriptions does not seem to contribute much in this experimental setup.

Nevertheless, our evaluation leaves open the question whether the relevance of different interaction modes may change over the course of development. For example, wrong utterances of a child are also frequently corrected by the mother. Children further strongly rely on these corrections during word learning. However, Chouinard and Clark (2003) showed that the frequency of parental corrections changes over time. Whereas they are often supplied at an early age, they become less frequent later on. At a later stage, other modes of interaction hence may become more important, e.g. the embedding of words into semantically rich object description (Weizman and Snow, 2001). It could be that a similar adaption in interaction patterns is advantageous for word learning by our computational model, too. This, however, remains to be validated in future experiments.

Another aspect that has not been investigate in this chapter is *active learning*, i.e. the active generation of word meaning hypotheses by the child. In our simulations, word meaning hypotheses were randomly sampled and subsequently corrected by the tutor. Other sampling strategies, however, may be more efficient. For example, a learner could choose such situations that lead to uncertain hypotheses (e.g. near word category boundaries) and request feedback for them. Active learning consequently would allow the learner to select those training samples for which he expects the highest gain in word knowledge. The incorporation of active learning strategies can consequently lead to an even more efficient word learning. Future research could validate this idea.

6

Summary

Language is a part of our organism and no less complicated than it.

Ludwig Wittgenstein (1889-1951)

In this thesis, I presented a computational framework for the acquisition of word meanings. Thereby, I pursued an interdisciplinary approach insofar as child development was taken as a role model for learning in artificial systems. The reason for this is that children exhibit excellent word learning capabilities which we ultimately would like to achieve in artificial systems, too. I therefore developed computational models that are inspired from findings in *developmental psychology* and *neurobiology*. In Chapter 2, I therefore presented a survey on current word learning theories as well as the neurobiological circuits that may underlie them. Based on the abundant literature I concluded that word learning is characterized by two distinct, but non-exclusive, paradigms: Firstly, it can be considered to be a mapping task, insofar as *cross-situational learning* can be used to find a mapping between word labels and word meanings. This paradigm inherently relies on the assumption that a learner has access to a number of potential word meanings that have been acquired prior to word learning. Even though *cross-situational learning* is known to be applied by children, it does not explain all observed learning patterns. For example, children do not possess a fully developed conceptual system onto which all novel words can be simply mapped. Potential word meanings in form of pre-established conceptual representations consequently do not always exist. For this reason, a second paradigm considers word learning to be equivalent to *concept formation*. There, the assumption is that concepts, i.e. potential word meanings, are not pre-established, but rather are constructed and gradually shaped through the use of language. Both paradigms are supported by findings from *developmental psychology*. In this thesis, I proposed computational models for each of them.

After having reviewed and categorized existing approaches according to the two paradigms, I first presented a model for unsupervised concept formation and word label mapping in Chapter 3. There, unsupervised learning was considered as one possible means for constructing potential word meanings prior to word learning. Of course, other mechanisms can serve the same purpose, but they were out of the scope of this thesis. In detail, the proposed system models two aspects: Firstly, how prior experience allows children to pre-conceptualize their environment in a data-driven way and, secondly, how word labels can be subsequently associated with the developed representations. A *dynamic neural field model*, that incorporates the principles of *Hebbian learning* and *homeostasis*, was implemented for these purposes. I thoroughly evaluated the model in various experiments, e.g. in a color naming scenario. There I showed that visual input alone drives the model to develop a topographic map in which colors are represented as a continuum. A subsequent incorporation of word labels, however, yields categories that link color names and their respective visual prototypes into unique representations.

Having introduced the model for unsupervised concept formation, I next discussed supervised word meaning acquisition, i.e. the process in which words guide the formation of corresponding conceptual representations, in Chapter 4. There, I argued that *Complementary Learning Systems (CLS) theory* provides one plausible explanation of the *fast mapping* and *slow mapping* processes that are typically observed in child development. The proposed computational model consequently is inspired by *CLS theory*. It comprises two complementary components which are specifically tailored to, first, rapidly memorize individual word-referent associations and, second, to decontextualize word meanings by abstracting common features among the referents of a word. Both components are recurrently coupled such that a gradual consolidation of the word knowledge is achieved. Following a systematic performance evaluation, the model was exemplarily applied in a visual scene description task in which words for the description of object relations were taught. Thereby, it was shown that the model results in learning patterns similar to those observed in child development.

Since the computational framework is inspired by findings on child development, it can be used to replicate findings, to test theories, or even to create new hypotheses regarding word learning by children. For example, it is known that children's word learning is strongly guided by learning constraints. The question, whether similar principles can be used to facilitate word meaning acquisition in the computational framework, thus, naturally arised. To provide an answer to this question, Chapter 5 showed on the example of the *mutual exclusivity principle* that learning constraints indeed help to acquire word meanings efficiently. In this scope, I additionally investigated the relevance of different interaction patterns that a tutor may use to teach the system. Thereby, it was experimentally shown that it is particularly important to correct wrong utterances of the learner. The evaluation hence provided important insights into how a tutor can facilitate the system's acquisition of word knowledge.

In summary, I presented a biologically inspired framework for word meaning acquisition. The framework comprises models for multiple aspects that seem to play a decisive role in children's word learning. Each part of the system was thoroughly evaluated. Thereby, the pursued interdisciplinary approach turned out to be promising with respect to both, understanding word learning by children more deeply as well as overcoming the difficulties of existing methods.

6.1. Conclusions

Regarding the four initial research goals, that were stated in Section 1.3, the following conclusions can be drawn.

Unveil the links between developmental psychology and neurobiology: Word learning theories often suggest that humans possess a dedicated system for the learning and representation of language. As I have reviewed in Chapter 2, however, *neurobiology* does not provide evidence in this respect. Findings rather demonstrate that the representations of words and their respective meanings are distributed all over the cortex, thereby recruiting areas whose primary functions are not language. Similarly, it is often assumed that the acquisition of word meanings exclusively relies on one learning principle. Here, I rather suggested that two principles may underlie the developmental patterns observed in children. The first one has been in the focus of Chapter 3, where I showed that a general associative model of cortical map formation is suitable to slowly acquire word representations via *Hebbian plasticity* and *homeostatic adaptation*. On the contrary, Chapter 4 investigated a principle for rapid word meaning acquisition. There, I argued that the *fast mapping* and *slow mapping* processes, which both constitute hallmarks of children's word learning, are based on complementary learning mechanisms in the hippocampus and neocortical sites. I reviewed a number of findings that support the suggested two-fold dissociation into a slow and a rapid learning process.

Provide biologically inspired computational models for word learning: For both identified learning processes, i.e. slow cortical and rapid hippocampal learning, I presented novel computational models. Thereby, the neurobiological circuits and principles which were hypothesized to underlie both learning processes strongly influenced the architectures and functions of the respective models. More precisely, I presented a *dynamic neural field (DNF)* for slow cortical map formation in Chapter 3. DNFs already had been successfully used in this domain, however, their ability to self-organize via learning had been very limited due to stability issues. In contrast to previous approaches, my model not only incorporates *Hebbian plasticity* to adapt network connectivity, but also relies on homeostatic principles being used in the central nervous system, namely *synaptic scaling* and *intrinsic plasticity*. These self-regulating principles counteract degradations in dynamic stability and therefore circumvent the problems existing methods are suffering from. As a model for hippocampal learning, I presented an adaptive *normalized Gaussian network (NGnet)* in Chapter 4. Similar to medial temporal lobe structures, my network possesses two different learning modes, i.e. a statistical learning based on Expectation-Maximization and a one-shot learning via hidden unit allocation. In accordance with biological memory consolidation based on the reactivation of hippocampal memories, I further recurrently coupled the network with an incremental discriminative feature extraction. This allows the model to gradually transfer acquired knowledge into the extracted features. In contrast to previous approaches, the model thus develops word meaning representations using adaptive feature spaces. This turned out to be of particular importance with respect to online learning. In Chapter 5, I additionally presented a model for the development of a *mutual exclusivity bias*. In contrast to the abovementioned networks, this model did not take inspiration from biology, but rather was solely based on a phenomenological description of the desired functionality. This is because it is currently unknown which circuits may underlie a similar kind of computation in the brain.

Evaluate the computational models thoroughly: In each chapter, the evaluation of the developed models pursued multiple objectives. Firstly, I applied them in simulations that emulated real biological experiments. This served the purpose of assessing whether the developed models exhibit computational characteristics similar to that of their neuronal role models. Thereby, I obtained results that are in accordance with findings from neurobiology and hence could prove the viability of the chosen implementations. Secondly, the models were applied in selected word learning scenarios. This not only allowed an investigation of the networks' computational characteristics, but also revealed whether the systems achieve a word learning performance as expected, i.e. qualitatively similar to that of children. Each model succeeded in this test, insofar as the results were as desired regarding both the dynamics of the learning process as well as the quality of the developed word representations. Where possible, I further applied the models in benchmark problems by which a performance comparison to state-of-art methods could be carried out. Overall, the networks performed very well, insofar as results superior to that of existing approaches were obtained.

Estimate the influence of human-robot interaction on word learning: In Chapter 2, I argued that – besides the two word learning processes – learning constraints play a pivotal role during children's acquisition of word meanings. One of these biases, namely the *mutual exclusivity principle*, was investigated in Chapter 5. It has previously been suggested that children are innately equipped with this principle. As I discussed, however, this seems problematic as the bias would hinder word learning in many cases where words are non-exclusively used. I therefore suggested that children develop a *mutual exclusivity principle* over the course of learning. To underpin this proposal, a computational model was presented that allows a learner to detect a mutually exclusive word use and hence circumvents the problems of an innate bias. The evaluation in Chapter 5, however, not only included learning biases that are internal to a child (e.g. the *mutual exclusivity principle*), but also those constraints that are externally applied by caregivers during social interaction with the learner. In fact, such social-pragmatic biases may even be more important, as caregivers can effectively constrain the learning environment (e.g. via single word use) and therefore influence which training samples a learner can use for word meaning acquisition. In detail, I investigated the effect of different conversation patterns on the system's word learning performance. A thorough evaluation revealed that different interaction modes are of varying importance in this respect. Thereby, corrections of wrong utterances of the learner were particularly relevant. The results thus highlighted that an active learning by the system is advantageous, as it allows the system to utter hypotheses on word meanings, request tutor's feedback to them, and finally may result in corrections of wrong hypotheses. Even though more experiments have to be carried out in future, the results of Chapter 5 thus gave important insights into how novel words can be most efficiently taught.

6.2. Suggestions for Future Research

In Chapters 3, 4, and 5 I already suggested future research directions regarding the individual computational models. This not only included possible algorithmic extensions of the models, but also experimental evaluations that could be carried out to validate

hypotheses regarding the functions of the networks. Here, I finally present ideas on an integrated computational framework for word meaning acquisition which combines the individual aspects that were investigated in this thesis. Such an integration was out of the scope of this thesis, but is necessary to finally equip robots with the desired word learning capabilities. Assembling a coherent overall framework hence may constitute the biggest challenge future research is facing.

Integrated Learning Architecture

An integration of the two proposed learning processes constitutes a first step towards such an overall framework. Even though both learning processes, i.e. rapid hippocampal and slow neocortical learning, exist in the brain and further seem to underlie word learning by children, it is unlikely that they run independently from each other. In fact, both mechanisms seem to be heavily intertwined, e.g. as suggested by *CLS theory* (McClelland et al., 1995). It is thus reasonable to combine the two computational models that were presented in Chapter 3 and Chapter 4. My proposal in this respect is illustrated in Fig. 6.1. In detail, I suggest to consider topographic map formation using our homeostatic DNF to be the standard learning paradigm. This means that slow associative learning underlies the formation of topographic map hierarchies which are learned based on the statistic of the sensory inputs presented to the system. This is in accordance with what has been shown in Chapter 3. In addition, I consider the rapid learning mechanism of the NGnet to be placed on top of such map hierarchies. In other words, activations within the maps may serve as input to the NGnet, where the different activity patterns can be rapidly memorized. The NGnet further can reactivate these memories, i.e. evoke activity patterns

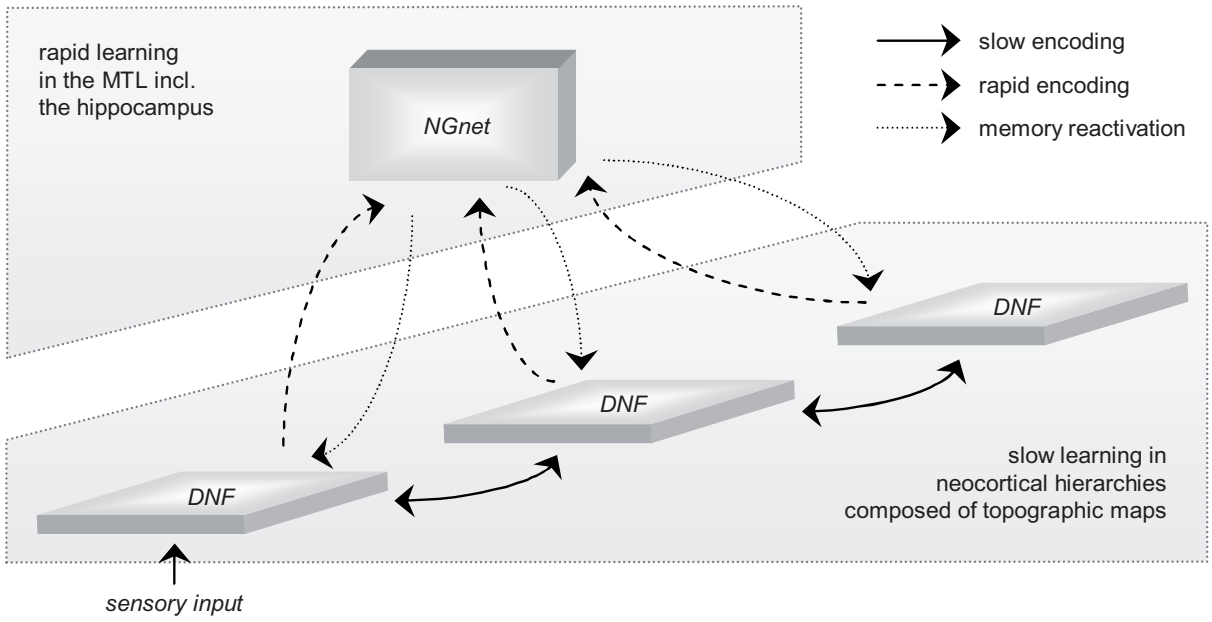


Figure 6.1.: Illustration of the proposed integrated learning architecture in which the rapid learning mechanism of the NGnet modulates the slow learning mechanism of the DNFs by altering the activity pattern statistic via reactivation.

in the topographic maps. It consequently is able to modulate the slow learning mechanism by altering the statistic of the activity patterns in the map. More precisely, the overall pattern statistic comprises those patterns that are evoked via externally supplied inputs plus those patterns that are internally reactivated by the NGnet. Depending on the ratio between reactivated and externally evoked patterns, map formation consequently can be dominated either by slow associative learning or the rapid learning mechanism of the NGnet. How to appropriately balance the two learning paradigms – or in other words, when to reactivate activity patterns and thereby drive map formation via the NGnet – hence constitutes an additional question future research has to answer.

Towards Incremental Vocabulary Growth

In each experiment presented in this thesis, word learning was limited to a small set of words. The meanings of these words further were acquired simultaneously and mainly independent from each other. An interaction between the learning of the different words only took place via the *mutual exclusivity principle*. In contrast, word learning by children is characterized by an incremental growth of the vocabulary. This means that children acquire their first words similar as it has been suggested in this thesis, i.e. starting with an empty lexicon. Afterwards, however, they already possess a certain amount of word knowledge which they can use during the learning of new words. In fact, it could be shown that already learned words help children to learn new words faster (Gershkoff-Stowe and Hahn, 2007). The exploitation of existing word knowledge hence may be one reason for what is called the *vocabulary spurt* – an exponential increase in the amount of acquired words which peaks around the age of 2 years (Ganger and Brent, 2004).

To model realistic word learning, an artificial system consequently should acquire words incrementally, thereby making use of already gained word knowledge. Future research hence should pursue the following two goals: Firstly, unveiling the mechanisms by which existing vocabulary may aid learning of novel words and, secondly, implementing those mechanisms in the computational framework. In the following, I just state three reasons why existing word knowledge can be beneficial.

- *Benefit from feature consolidation:*
  As it has been shown in Section 4.2.3, learning extracts word meaning relevant features for each word. For example, the relative distance between two objects is a relevant feature for words describing spatial object relations (e.g. *is above*). During incremental vocabulary acquisition, it may be beneficial to re-use such features, i.e. to gradually extend the pool of feature dimensions in which the meaning of a novel word can be grounded. For example, a relative object distance could also be relevant to the meaning of *overlapping*. This idea is supported by a study of Hoffman et al. (2008) who showed that already acquired feature knowledge helps children to learn new words. Overall, incorporating a feature consolidation may be beneficial as individual feature dimensions can be relevant for multiple words and hence should be re-used rather than being extracted multiple times.
- *Benefit from attentional pruning of word meaning hypotheses:*
  Consider a mother who presents a red ball to its child and says *'Look! This one is not blue, it's red.'* Furthermore, assume that the child already knows the meaning

of *blue*, whereas *red* constitutes a novel word. The child could infer multiple aspects from the mother's utterance. This includes that the uses of *blue* and *red* seem to be mutually exclusive (see Section 5.2.1), but also that both words stem from the same domain (color). In detail, given the child's knowledge that *blue* refers to a color, it can infer that *red* also refers to a color. The child consequently can focus attention on the color dimensions, thereby pruning many wrong word meaning hypotheses.

- *Benefit from explicit contrast:*
  In the previous example, the mother explicitly contrasted the meanings of *blue* and *red*. The child thus knows that both words refer to distinct things. Since the child already learned a category representation for the meaning of *blue*, it knows which region of the input space cannot be covered by a *red* category. Explicit contrasting hence not only focuses attention to relevant feature dimensions, but also constrains the potential word meanings within the relevant feature space. This principle not only holds for the learning of novel words, of course. Already existing word knowledge could be consolidated via a similar mechanism, i.e. given knowledge on an exclusive word use, the respective internal word categories could be adapted such that they do not overlap anymore.

Provide Insights on Child-like Learning Processes to Developmental Psychology and Neurobiology

The interdisciplinary approach, that was pursued in this thesis, established a link between the individual disciplines *developmental psychology*, *neurobiology*, and *computer science*. More precisely, I stated commonalities between findings from *developmental psychology* and *neurobiology* based on which I developed biologically inspired computational models for word learning. As it was illustrated in Fig. 1.3, however, it is desirable not only to consider an unidirectional influence, insofar as computational modeling takes inspiration from the other two disciplines. Future research rather should aim at closing the loop, such that *developmental psychology* and *neurobiology* can gain insights from our computational models, too. The ultimate goal of our future research therefore is to develop a computational framework based on which novel hypotheses regarding word learning in children can be stated and finally validated in real experiments with the help of the other two disciplines.

A

Clustering Multivariate Normal Distributions

The clustering of multivariate normal distributions is of interest in many domains. Its main technical application is speech recognition using *Hidden Markov Models (HMMs)*. Thereby, the aim is to tie acoustically similar states which each use Gaussian mixtures to represent the observed speech data (Young and Woodland, 1993). In the present work, the clustering of Gaussians plays a key role in the merging of the hidden units of an NGnet (cf. Section 4.2.2). This appendix gives a detailed derivation of the formulas that are used during this process.

In the following, the general case of clustering K multivariate Gaussian distributions $G(\boldsymbol{a}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with $i = 1, \ldots, K$ is considered. Thereby, a Gaussian distribution

$$G(\boldsymbol{a}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\mathcal{D}/2}|\boldsymbol{\Sigma}_i|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_i)\right) \tag{A.1}$$

is parametrized by its mean $\boldsymbol{\mu_i}$ and its covariance matrix $\boldsymbol{\Sigma_i}$. Furthermore, $\mathcal{D}$ denotes the dimensionality of the space, i.e. $\boldsymbol{a} \in \mathbb{R}^{\mathcal{D}}$. The quality of a clustering solution can be measured by the functional $\mathcal{F}$

$$\mathcal{F} = \sum_{i=1}^{K} \omega_i \cdot D(\, G(\boldsymbol{a}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \,||\, G(\boldsymbol{a}, \boldsymbol{m}, \boldsymbol{S}) \,), \tag{A.2}$$

where $G(\boldsymbol{a}, \boldsymbol{m}, \boldsymbol{S})$ denotes the resulting Gaussian, D is a divergence measure, and ω_i are weights that control the influence of the individual distributions on the clustering process. W.l.o.g. it is assumed that $\sum_i \omega_i = 1$. We consequently aim at finding an optimal Gaussian $G(\boldsymbol{a}, \boldsymbol{m}^*, \boldsymbol{S}^*)$ that minimizes $\mathcal{F}$.

Existing approaches for solving this task differ in the divergence measures they use. In the following, two of them – namely *Kullback-Leibler divergence* based clustering (Davis and Dhillon, 2006) and *Jenson-Shannon divergence* based clustering (Myrvoll and Soong, 2003) – are reviewed.

Kullback-Leibler Divergence Based Clustering

The *Kullback-Leibler divergence* is defined as

$$D_{KL}(\, p(\boldsymbol{a}) \,||\, q(\boldsymbol{a}) \,) = \int_{\boldsymbol{a}} p(\boldsymbol{a}) \cdot \log \frac{p(\boldsymbol{a})}{q(\boldsymbol{a})} \, d\boldsymbol{a}, \tag{A.3}$$

which for Gaussian pdfs simplifies to

$$\begin{aligned} D_{KL}(\, G(\boldsymbol{a}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \,||\, G(\boldsymbol{a}, \boldsymbol{m}, \boldsymbol{S}) \,) &= \frac{1}{2} \cdot \big(\mathrm{trace}(\boldsymbol{\Sigma}_i \boldsymbol{S}^{-1}) - \log |\boldsymbol{\Sigma}_i \boldsymbol{S}^{-1}| - d \big) \\ &\quad + \frac{1}{2} \cdot (\boldsymbol{\mu}_i - \boldsymbol{m})^T \boldsymbol{S}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{m}) \\ &= \frac{1}{2} \cdot \left[D_{Burg}(\boldsymbol{\Sigma}_i, \boldsymbol{S}) + D_{Mahal}(\boldsymbol{\mu}_i, \boldsymbol{m}, \boldsymbol{S}) \right]. \end{aligned} \tag{A.4}$$

Thereby, D_{Burg} denotes the *Burg matrix divergence* and D_{Mahal} the *Mahalanobis distance*. When using this divergence measure the optimization function in Eq. (A.2) becomes

$$\mathcal{F}_{KL} = \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[D_{Burg}(\boldsymbol{\Sigma}_i, \boldsymbol{S}) + D_{Mahal}(\boldsymbol{\mu}_i, \boldsymbol{m}, \boldsymbol{S}) \right]. \tag{A.5}$$

The mean $\boldsymbol{m}^*$ and the covariance matrix $\boldsymbol{S}^*$ of the optimal Gaussian $G(\boldsymbol{a}, \boldsymbol{m}^*, \boldsymbol{S}^*)$ can be obtained by setting the derivative of $\mathcal{F}_{KL}$ with respect to the parameters to 0.

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial \mathcal{F}_{KL}}{\partial \boldsymbol{m}} \\ &= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[\frac{\partial D_{Burg}(\boldsymbol{\Sigma}_i, \boldsymbol{S})}{\partial \boldsymbol{m}} + \frac{\partial D_{Mahal}(\boldsymbol{\mu}_i, \boldsymbol{m}, \boldsymbol{S})}{\partial \boldsymbol{m}} \right] \\ &= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[0 + 2 \cdot \boldsymbol{S}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{m}) \right] \\ &= \boldsymbol{S}^{-1} \cdot \sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\mu}_i - \boldsymbol{m}) \\ &= -\boldsymbol{m} + \sum_{i=1}^{K} \omega_i \cdot \boldsymbol{\mu}_i. \end{aligned} \tag{A.6}$$

Consequently, the mean of the optimal Gaussian is given by

$$\boldsymbol{m}^* = \sum_{i=1}^{K} \omega_i \cdot \boldsymbol{\mu}_i. \tag{A.7}$$

In a similar way, the covariance matrix can be calculated by setting

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial \mathcal{F}_{KL}}{\partial \boldsymbol{S}^{-1}} \\ &= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[\frac{\partial D_{Burg}(\boldsymbol{\Sigma}_i, \boldsymbol{S})}{\partial \boldsymbol{S}^{-1}} + \frac{\partial D_{Mahal}(\boldsymbol{\mu}_i, \boldsymbol{m}, \boldsymbol{S})}{\partial \boldsymbol{S}^{-1}} \right] \\ &= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[\boldsymbol{\Sigma}_i^T - \frac{1}{|\boldsymbol{\Sigma}_i \boldsymbol{S}|} \cdot |\boldsymbol{\Sigma}_i \boldsymbol{S}| \cdot \boldsymbol{S}^T + (\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T \right] \\ &= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[\boldsymbol{\Sigma}_i - \boldsymbol{S} + (\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T \right] \\ &= -\boldsymbol{S} + \sum_{i=1}^{K} \omega_i \cdot \left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T \right], \end{aligned} \tag{A.8}$$

which leads to

$$\boldsymbol{S}^* = \sum_{i=1}^{K} \omega_i \cdot \left(\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{m}^*)(\boldsymbol{\mu}_i - \boldsymbol{m}^*)^T \right). \tag{A.9}$$

Jenson-Shannon Divergence Based Clustering

The *Jenson-Shannon divergence* is a symmetrical version of the *Kullback-Leibler divergence* and is defined as

$$D_{JS}(\, p(\boldsymbol{a}) \,||\, q(\boldsymbol{a}) \,) = \frac{1}{2} \left(D_{KL}(\, p(\boldsymbol{a}) \,||\, q(\boldsymbol{a}) \,) + D_{KL}(\, q(\boldsymbol{a}) \,||\, p(\boldsymbol{a}) \,) \right). \tag{A.10}$$

For Gaussian pdfs it simplifies to

$$\begin{aligned} D_{JS}(\, G(\boldsymbol{a}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \,||\, G(\boldsymbol{a}, \boldsymbol{m}, \boldsymbol{S}) \,) &= \frac{1}{4} \cdot \operatorname{trace} \left\{ (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T \right. \\ &\quad \left. + \boldsymbol{\Sigma}_i \boldsymbol{S}^{-1} + \boldsymbol{S} \boldsymbol{\Sigma}_i^{-1} - 2\boldsymbol{I} \right\}. \end{aligned} \tag{A.11}$$

Using D_{JS} in Eq. (A.2) the function to be optimized becomes

$$\mathcal{F}_{JS} = \frac{1}{4} \cdot \sum_{i=1}^{K} \omega_i \cdot \operatorname{trace} \left\{ (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T + \boldsymbol{\Sigma}_i \boldsymbol{S}^{-1} + \boldsymbol{S} \boldsymbol{\Sigma}_i^{-1} - 2\boldsymbol{I} \right\}. \tag{A.12}$$

As before, the parameters $\boldsymbol{m}^*$ and $\boldsymbol{S}^*$ of the optimal Gaussian can be obtained by setting the derivative of $\mathcal{F}_{JS}$ with respect to the parameters to 0.

$$\begin{aligned}
0 &\stackrel{!}{=} \frac{\partial \mathcal{F}}{\partial \boldsymbol{m}} \\
&= \frac{1}{4} \cdot \sum_{i=1}^{K} \omega_i \cdot \frac{\partial \text{trace}\left\{(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T\right\}}{\partial(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T} \cdot \frac{\partial(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T}{\partial \boldsymbol{m}} \\
&= \frac{1}{4} \cdot \sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})^T \cdot 2 \cdot (\boldsymbol{\mu}_i - \boldsymbol{m}) \\
&= \frac{1}{2} \cdot \sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\mu}_i - \boldsymbol{m}) \\
&= -\sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})\boldsymbol{m} + \sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})\boldsymbol{\mu}_i. \qquad \text{(A.13)}
\end{aligned}$$

Consequently, the mean of the optimal Gaussian can be calculated by

$$\boldsymbol{m}^* = \left[\sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{*^{-1}})\right]^{-1} * \left[\sum_{i=1}^{K} \omega_i \cdot (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{*^{-1}})\boldsymbol{\mu}_i\right]. \qquad \text{(A.14)}$$

Similarly, it is

$$\begin{aligned}
0 &\stackrel{!}{=} \frac{\partial \mathcal{F}_{JS}}{\partial \boldsymbol{S}^{-1}} \\
&= \frac{1}{4} \cdot \sum_{i=1}^{K} \omega_i \cdot \frac{\partial \text{trace}\left\{(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{S}^{-1})(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T + \boldsymbol{\Sigma}_i\boldsymbol{S}^{-1} + \boldsymbol{S}\boldsymbol{\Sigma}_i^{-1} - 2\boldsymbol{I}\right\}}{\partial \boldsymbol{S}^{-1}} \\
&= \frac{1}{4} \cdot \sum_{i=1}^{K} \omega_i \cdot \left[(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T + \boldsymbol{\Sigma}_i - \boldsymbol{S}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{S}\right] \\
&= \left\{\sum_{i=1}^{K} \omega_i \cdot \left[(\boldsymbol{\mu}_i - \boldsymbol{m})(\boldsymbol{\mu}_i - \boldsymbol{m})^T + \boldsymbol{\Sigma}_i\right]\right\} - \boldsymbol{S}\left\{\sum_{i=1}^{K} \omega_i \cdot \boldsymbol{\Sigma}_i^{-1}\right\}\boldsymbol{S}, \qquad \text{(A.15)}
\end{aligned}$$

which fits the matrix Ricatti equation

$$0 = \boldsymbol{A} + \boldsymbol{B}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{B}^* - \boldsymbol{S}\boldsymbol{C}\boldsymbol{S} \qquad \text{(A.16)}$$

with

$$\begin{aligned}
\boldsymbol{A} &= \sum_{i=1}^{K} \omega_i \cdot \left[(\boldsymbol{\mu}_i - \boldsymbol{m}^*)(\boldsymbol{\mu}_i - \boldsymbol{m}^*)^T + \boldsymbol{\Sigma}_i\right] \qquad \text{(A.17)} \\
\boldsymbol{B} &= \boldsymbol{0} \qquad \text{(A.18)} \\
\boldsymbol{C} &= \sum_{i=1}^{K} \omega_i \cdot \boldsymbol{\Sigma}_i^{-1}. \qquad \text{(A.19)}
\end{aligned}$$

A solution to this equation is given by

$$\boldsymbol{S}^* = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] * [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d]^{-1} \tag{A.20}$$

with $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$ being the upper and lower parts of vector $\boldsymbol{v}_i$

$$\boldsymbol{v}_i = \begin{bmatrix} \boldsymbol{u}_i \\ \boldsymbol{w}_i \end{bmatrix}, \tag{A.21}$$

where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ are the eigenvectors corresponding to the d positive eigenvalues (sorted in descending order) of matrix

$$\begin{bmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{C} & -\boldsymbol{B}^* \end{bmatrix}. \tag{A.22}$$

Since the formulas for the optimal mean $\boldsymbol{m}^*$ and covariance matrix $\boldsymbol{S}^*$ depend on each other, their calculation involves the iterative application of Eq. (A.14) and Eq. (A.20). Thereby, the optimal mean of the Kullback-Leibler divergence based clustering, i.e.

$$\boldsymbol{m}^* = \sum_{i=1}^{K} \omega_i \cdot \boldsymbol{\mu}_i, \tag{A.23}$$

constitutes a good starting point of this iterative process.

B

Information-Theoretic Feature Extraction

The aim of a feature extraction is to find a function f that transforms an input pattern $\boldsymbol{x}$ into a feature pattern $\boldsymbol{y}$, i.e. $\boldsymbol{y} = f(\boldsymbol{x})$, such that the feature patterns are better suited for a subsequent task than the corresponding input patterns are. With respect to a classification task, the primary goal of a feature extraction consequently is to find feature dimensions that facilitate an association of the patterns $\boldsymbol{y}$ with their respective class labels c. Multiple approaches for a class-discriminative feature extraction exist. Here, an information-theoretic approach recently proposed by Hild et al. (2006) is of particular interest as it constitutes one of the components of the computational model presented in Section 4.2. This appendix entails a detailed review of the method of Hild et al. and further provides a derivation of the important formulas.

Mutual Information Criterion

To learn a feature extraction function f, a criterion for the quality of the mapping $f : \mathcal{S}_{\boldsymbol{x}} \mapsto \mathcal{S}_{\boldsymbol{y}}$ has to be defined. Here, $\mathcal{S}_{\boldsymbol{x}}$ denotes the input space, whereas $\mathcal{S}_{\boldsymbol{y}}$ refers to the feature space. An information-theoretic criterion suited for a classification task is the mutual information between the feature patterns and the class labels. It is defined as

$$I(\boldsymbol{Y}; C) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|C), \tag{B.1}$$

where $H(\boldsymbol{Y})$ and $H(\boldsymbol{Y}|C)$ denote Shannon's marginal entropy and conditional entropy, respectively. The mutual information describes the amount of information that the feature patterns $\boldsymbol{y}$ carry about the class labels c. For the extraction of class-discriminative features, we consequently strive for a function f that maximizes $I(\boldsymbol{Y}; C)$.

The mutual information can be calculated according to

$$\begin{aligned} I(\boldsymbol{Y};C) &= -\int_{\boldsymbol{y}} p(\boldsymbol{y}) \cdot \log\left[p(\boldsymbol{y})\right] \mathrm{d}\boldsymbol{y} \; + \; \sum_c p(c) \cdot \int_{\boldsymbol{y}} p(\boldsymbol{y}|c) \cdot \log\left[p(\boldsymbol{y}|c)\right] \mathrm{d}\boldsymbol{y} \\ &= -\int_{\boldsymbol{y}} \sum_c p(\boldsymbol{y},c) \cdot \log\left[p(\boldsymbol{y})\right] \mathrm{d}\boldsymbol{y} \; + \; \sum_c \frac{p(\boldsymbol{y},c)}{p(\boldsymbol{y}|c)} \cdot \int_{\boldsymbol{y}} p(\boldsymbol{y}|c) \cdot \log\left[\frac{p(\boldsymbol{y},c)}{p(c)}\right] \mathrm{d}\boldsymbol{y} \\ &= \sum_c \int_{\boldsymbol{y}} p(\boldsymbol{y},c) \cdot \log\left[\frac{p(\boldsymbol{y},c)}{p(\boldsymbol{y}) \cdot p(c)}\right] \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{B.2}$$

Here, $p(\boldsymbol{y})$ and $p(c)$ are the marginal probabilities, $p(\boldsymbol{y}|c)$ is the conditional probability, and $p(\boldsymbol{y},c)$ is the joint probability between the feature patterns and the class labels. The problem in using this criterion is that the aforementioned probabilities are typically unknown. They rather have to be estimated via computationally expensive approaches. Hild et al. (2006) suggested to reduce this burden by using Renyi's quadratic entropy (Renyi, 1970) instead of Shannon's entropy. Renyi's quadratic entropy is defined as

$$\begin{aligned} H_2(\boldsymbol{Y}) &= -\log \int_{\boldsymbol{y}} p(\boldsymbol{y})^2 \mathrm{d}\boldsymbol{y} \\ H_2(\boldsymbol{Y}|C) &= -\sum_c p(c) \cdot \log \int_{\boldsymbol{y}} p(\boldsymbol{y}|c)^2 \mathrm{d}\boldsymbol{y}, \end{aligned} \tag{B.3}$$

which leads to

$$\begin{aligned} I_2(\boldsymbol{Y};C) &= H_2(\boldsymbol{Y}) - H_2(\boldsymbol{Y}|C) \\ &= -\log \int_{\boldsymbol{y}} p(\boldsymbol{y})^2 \mathrm{d}\boldsymbol{y} + \sum_c p(c) \cdot \log \int_{\boldsymbol{y}} p(\boldsymbol{y}|c)^2 \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{B.4}$$

Hild et al. further proposed to use Parzen window density estimation to approximate the required probabilities on a per sample basis. Parzen windowing (Parzen, 1962) therefore places a kernel at each sample. The sum of these kernels finally yields the sample density. We use Gaussian kernels of the form

$$G(\boldsymbol{y}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\mathcal{D}_y/2} |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \boldsymbol{y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\right), \tag{B.5}$$

by which the probability $p(\boldsymbol{y})$ can be approximated as

$$p(\boldsymbol{y}) = \frac{1}{N} \cdot \sum_{i=1}^{N} G(\boldsymbol{y} - \boldsymbol{y}_i, \sigma \boldsymbol{I}). \tag{B.6}$$

Thereby, $\boldsymbol{y}_i$ with $i = 1 \ldots N$ are the individual samples at which the kernels have been placed and σ is the standard deviation of the kernels. The strength of combining Parzen windowing with Renyi's quadratic entropy measure arises from the fact that

$$\int_{\boldsymbol{y}} \left[G(\boldsymbol{y} - \boldsymbol{y}_i, \boldsymbol{\Sigma}_i) \cdot G(\boldsymbol{y} - \boldsymbol{y}_j, \boldsymbol{\Sigma}_j)\right] \mathrm{d}\boldsymbol{y} = G(\boldsymbol{y}_i - \boldsymbol{y}_j, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j). \tag{B.7}$$

This means that the convolution of two Gaussians centered at individual samples $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ can be calculated by evaluating one Gaussian centered at $\boldsymbol{y}_j$ at the point $\boldsymbol{y}_i$. Thereby, the covariance matrix of that Gaussian equals the sum of the covariance matrices of the individual Gaussians. By using Eq. (B.7) the mutual information criterion of Eq. (B.4) can be calculated as

$$\begin{aligned} I_2(\boldsymbol{Y};C) &= -\log \int_{\boldsymbol{y}} \left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G(\boldsymbol{y}-\boldsymbol{y}_i, \sigma\boldsymbol{I}) \cdot G(\boldsymbol{y}-\boldsymbol{y}_j, \sigma\boldsymbol{I}) \right) \mathrm{d}\boldsymbol{y} \\ &\quad + \sum_{c=1}^{K} \frac{N_c}{N} \cdot \log \int_{\boldsymbol{y}} \left(\frac{1}{N_c^2} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} G(\boldsymbol{y}-\boldsymbol{y}_i^{(c)}, \sigma\boldsymbol{I}) \cdot G(\boldsymbol{y}-\boldsymbol{y}_j^{(c)}, \sigma\boldsymbol{I}) \right) \mathrm{d}\boldsymbol{y} \\ &= -\log \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G(\boldsymbol{y}_i-\boldsymbol{y}_j, 2\sigma\boldsymbol{I}) \\ &\quad + \sum_{c=1}^{K} \frac{N_c}{N} \cdot \log \frac{1}{N_c^2} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} G(\boldsymbol{y}_i^{(c)}-\boldsymbol{y}_j^{(c)}, 2\sigma\boldsymbol{I}). \end{aligned} \tag{B.8}$$

Here, $p(c)$ is set according to $p(c) = N_c/N$, where N_c denotes the number of samples in class c, K is the number of classes, and $N = \sum_{c=1}^{K} N_c$ is the total number of samples. Furthermore, $\boldsymbol{y}_i^{(c)}$ refers to the i-th sample of class c, whereas $\boldsymbol{y}_i$ refers to the i-th sample of the overall training set.

Derivatives for Learning

Finding a function $\boldsymbol{y} = f(\boldsymbol{x})$ that maximizes Eq. (B.8) hence corresponds to learning discriminative features. Given that the derivative $\partial\boldsymbol{y}/\partial f$ exists, this can be achieved via stochastic gradient ascent on the mutual information criterion. More precisely, for a linear feature extraction of form $\boldsymbol{y} = \boldsymbol{R} \cdot \boldsymbol{x}$ (like the one applied in Section 4.2.3), the feature extraction matrix $\boldsymbol{R}$ can be iteratively learned according to

$$\begin{aligned} \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t + \eta \cdot \frac{\partial I_2(\boldsymbol{Y};C)}{\partial \boldsymbol{R}_t} \\ &= \boldsymbol{R}_t + \eta \cdot \left(\frac{\partial H_2(\boldsymbol{Y})}{\partial \boldsymbol{R}_t} - \frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \boldsymbol{R}_t} \right). \end{aligned} \tag{B.9}$$

Here, η is a learning rate. Due to the fact that

$$\frac{\partial \boldsymbol{y}_k}{\partial \boldsymbol{R}} = \boldsymbol{x}_k^T, \tag{B.10}$$

the derivative of the mutual information criterion with respect to the feature extraction matrix $\boldsymbol{R}$ (see Eq. (B.9)) can be calculated on the basis of individual training samples $(\boldsymbol{x}_t, c_t)$, where $\boldsymbol{x}_t$ refers to the t-th input pattern and c_t is the associated class label.

Thereby, the following formulas are used:

$$\begin{aligned}
\frac{\partial H_2(\boldsymbol{Y})}{\partial \boldsymbol{R}} &= \sum_{k=1}^{N} \frac{\partial H_2(\boldsymbol{Y})}{\partial \boldsymbol{y}_k} \cdot \frac{\partial \boldsymbol{y}_k}{\partial \boldsymbol{R}} \\
&= \sum_{k=1}^{N} \Bigg\{ -\frac{1}{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma\boldsymbol{I})} \cdot \frac{1}{N^2} \\
&\qquad \cdot \Bigg(\sum_{l=1}^{N} \left[G(\boldsymbol{y}_l - \boldsymbol{y}_k, 2\sigma\boldsymbol{I}) \cdot (-\frac{1}{2}) \cdot \frac{1}{4\sigma^2} \cdot (\boldsymbol{y}_l - \boldsymbol{y}_k) \cdot (-1) \right] \\
&\qquad\quad + \sum_{l=1}^{N} \left[G(\boldsymbol{y}_k - \boldsymbol{y}_l, 2\sigma\boldsymbol{I}) \cdot (-\frac{1}{2}) \cdot \frac{1}{4\sigma^2} \cdot (\boldsymbol{y}_k - \boldsymbol{y}_l) \right] \Bigg) \\
&\qquad \cdot \frac{\partial \boldsymbol{y}_k}{\partial \boldsymbol{R}} \Bigg\} \\
&= \sum_{k=1}^{N}\sum_{l=1}^{N} \frac{G(\boldsymbol{y}_l - \boldsymbol{y}_k, 2\sigma\boldsymbol{I}) \cdot (\boldsymbol{y}_k - \boldsymbol{y}_l)}{4\sigma \cdot \sum_{i=1}^{N}\sum_{j=1}^{N} G(\boldsymbol{y}_i - \boldsymbol{y}_j, 2\sigma\boldsymbol{I})} \cdot \frac{\partial \boldsymbol{y}_k}{\partial \boldsymbol{R}}
\end{aligned} \tag{B.11}$$

$$\begin{aligned}
-\frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \boldsymbol{R}} &= -\sum_{c=1}^{K}\sum_{k=1}^{N_c} \frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \boldsymbol{y}_k^{(c)}} \cdot \frac{\partial \boldsymbol{y}_k^{(c)}}{\partial \boldsymbol{R}} \\
&= \sum_{c=1}^{K}\sum_{k=1}^{N_c} \Bigg\{ \frac{N_c}{N} \cdot \frac{1}{\frac{1}{N_c^2}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c} G(\boldsymbol{y}_i^{(c)} - \boldsymbol{y}_j^{(c)}, 2\sigma\boldsymbol{I})} \cdot \frac{1}{N_c^2} \\
&\qquad \cdot \Bigg(\sum_{l=1}^{N_c} \left[G(\boldsymbol{y}_l^{(c)} - \boldsymbol{y}_k^{(c)}, 2\sigma\boldsymbol{I}) \cdot (-\frac{1}{2}) \cdot \frac{1}{4\sigma^2} \cdot (\boldsymbol{y}_l^{(c)} - \boldsymbol{y}_k^{(c)}) \cdot (-1) \right] \\
&\qquad\quad + \sum_{l=1}^{N_c} \left[G(\boldsymbol{y}_k^{(c)} - \boldsymbol{y}_l^{(c)}, 2\sigma\boldsymbol{I}) \cdot (-\frac{1}{2}) \cdot \frac{1}{4\sigma^2} \cdot (\boldsymbol{y}_k^{(c)} - \boldsymbol{y}_l^{(c)}) \right] \Bigg) \\
&\qquad \cdot \frac{\partial \boldsymbol{y}_k^{(c)}}{\partial \boldsymbol{R}} \Bigg\} \\
&= \sum_{c=1}^{K}\sum_{k=1}^{N_c}\sum_{l=1}^{N_c} \frac{N_c \cdot G(\boldsymbol{y}_l^{(c)} - \boldsymbol{y}_k^{(c)}, 2\sigma\boldsymbol{I}) \cdot (\boldsymbol{y}_l^{(c)} - \boldsymbol{y}_k^{(c)})}{4\sigma N \cdot \sum_{i=1}^{N_c}\sum_{j=1}^{N_c} G(\boldsymbol{y}_i^{(c)} - \boldsymbol{y}_j^{(c)}, 2\sigma\boldsymbol{I})} \cdot \frac{\partial \boldsymbol{y}_k^{(c)}}{\partial \boldsymbol{R}}
\end{aligned} \tag{B.12}$$

As can be seen, the maximization of the mutual information criterion results in pair-wise interactions between the feature patterns. These interactions can be interpreted in terms of *information forces* (Torkkola, 2003) that the different samples exert on each other. According to

$$\frac{\partial H_2(\boldsymbol{Y})}{\partial \boldsymbol{y}_k} \propto \sum_{l=1}^{N} \gamma_{kl} \cdot (\boldsymbol{y}_k - \boldsymbol{y}_l) \tag{B.13}$$

the first derivative results in repulsion forces that push the feature pattern $\boldsymbol{y}_k$ away from all other patterns of the training set. Thereby, γ_{kl} is a factor that influences the strength

of the force between $\boldsymbol{y}_k$ and $\boldsymbol{y}_l$. The factors γ depend on the pair-wise distances between the feature patterns. In contrast to the repulsion forces caused by $\partial H_2(\boldsymbol{Y})/\partial \boldsymbol{y}_k$, the derivative

$$-\frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \boldsymbol{y}_k^{(c)}} \propto \sum_{l=1}^{N_c} \gamma_{kl}^{(c)} \cdot (\boldsymbol{y}_l^{(c)} - \boldsymbol{y}_k^{(c)}) \tag{B.14}$$

induces attraction forces between samples of the same class. The mutual information criterion is consequently maximized when the feature extraction matrix $\boldsymbol{R}$ transforms inputs of the same class into feature patterns that are close to each other. At the same time, however, the distance to the feature patterns of other classes is maximized. This finally results in a feature space in which the different patterns cluster according to the classes they belong to. It is worth noting that other methods for discriminative feature extraction apply in part similar techniques. Linear Discriminant Analysis (LDA), for example, serves the maximization of the ratio of the inter-class variance and the intra-class variance.

Sequential Learning Scheme

The pair-wise interactions result in an algorithmic complexity of $O(N^2)$, i.e. the computational cost quickly increases as the number of training samples increases. Since mutual information based feature extraction is a statistical learning method, a large training set is needed. For this reason, it is important to reduce the algorithmic complexity. According to Hild et al. (2006), the present learning method can be approximated by an algorithm of complexity $O(N)$. Key to this approach is that each sample $\boldsymbol{y}_k$ does not have to interact with all other samples $\boldsymbol{y}_l$ with $l = 1 \ldots N$, but only with a randomly chosen sample $\boldsymbol{y}_l$ with $l \in \{1, \ldots, N\}$. W.l.o.g. we assume that a sample $\boldsymbol{y}_k$ only interacts with sample $\boldsymbol{y}_{k+1}$. It can be shown, that this simplified interaction converges in the limit to the result of the original algorithm, if the training samples are presented multiple times and the sample order is randomized for each presentation (Erdogmus et al., 2003).

Accordingly, the mutual information criterion of Eq. (B.8) simplifies to

$$\begin{aligned} I_2(\boldsymbol{Y}; C) &= -\log \frac{1}{N} \sum_{i=1}^{N} G(\boldsymbol{y}_i - \boldsymbol{y}_{i+1}, 2\sigma \boldsymbol{I}) \\ &\quad + \sum_{c=1}^{K} \frac{N_c}{N} \cdot \log \frac{1}{N_c} \sum_{i=1}^{N_c} G(\boldsymbol{y}_i^{(c)} - \boldsymbol{y}_{i+1}^{(c)}, 2\sigma \boldsymbol{I}). \end{aligned} \tag{B.15}$$

If we set $\triangle \boldsymbol{y}_k = \boldsymbol{y}_k - \boldsymbol{y}_{k+1}$ and $\triangle \boldsymbol{x}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k+1}$, then the respective derivatives for learning the feature extraction matrix $\boldsymbol{R}$ are:

$$\frac{\partial \triangle \boldsymbol{y}_k}{\partial \boldsymbol{R}} = \triangle \boldsymbol{x}_k^T \tag{B.16}$$

$$\begin{aligned} \frac{\partial H_2(\boldsymbol{Y})}{\partial \boldsymbol{R}} &= \sum_{k=1}^{N} \frac{\partial H_2(\boldsymbol{Y})}{\partial \triangle \boldsymbol{y}_k} \cdot \frac{\partial \triangle \boldsymbol{y}_k}{\partial \boldsymbol{R}} \\ &= \sum_{k=1}^{N} \frac{G(\triangle \boldsymbol{y}_k, 2\sigma \boldsymbol{I})}{8\sigma \cdot \sum_{i=1}^{N} G(\triangle \boldsymbol{y}_i, 2\sigma \boldsymbol{I})} \cdot \triangle \boldsymbol{y}_k \cdot \triangle \boldsymbol{x}_k^T \end{aligned} \tag{B.17}$$

$$
\begin{aligned}
-\frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \boldsymbol{R}} &= -\sum_{c=1}^{K}\sum_{k=1}^{N_c}\frac{\partial H_2(\boldsymbol{Y}|C)}{\partial \triangle \boldsymbol{y}_k^{(c)}}\cdot\frac{\partial \triangle \boldsymbol{y}_k^{(c)}}{\partial \boldsymbol{R}} \\
&= -\sum_{c=1}^{K}\sum_{k=1}^{N_c}\frac{N_c \cdot G(\triangle \boldsymbol{y}_k^{(c)}, 2\sigma \boldsymbol{I})}{8\sigma N \cdot \sum_{i=1}^{N_c} G(\triangle \boldsymbol{y}_i^{(c)}, 2\sigma \boldsymbol{I})}\cdot \triangle \boldsymbol{y}_k^{(c)} \cdot \triangle \boldsymbol{x}_k^{(c)^T}. \quad \text{(B.18)}
\end{aligned}
$$

In summary, the feature extraction matrix $\boldsymbol{R}$ can be learned as shown in Algorithm B.1. Given a training set composed of samples $(\boldsymbol{x}_t, c_t)$, the method iteratively updates $\boldsymbol{R}$ according to the information forces the samples exert on each other. $\boldsymbol{R}$ finally produces a class-discriminative feature space in which the patterns of the different classes are separated, whereas those of the same class are close-by.

Algorithm B.1 Sequential Feature Extraction

{Inputs}
$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$
$C = \{c_1, \ldots, c_N\}$
$\boldsymbol{R} = \boldsymbol{R}_{INIT}$

{Create Class-Specific Input Sets}
$\boldsymbol{X}^{(c)} \leftarrow \{\boldsymbol{x}_1^{(c)}, \ldots, \boldsymbol{x}_{N_c}^{(c)}\}$ with $c = 1, \ldots, M$

loop

{Change Presentation Order}
Randomize the elements of $\boldsymbol{X}$ and $\boldsymbol{X}^{(c)}$, $\forall c$

{Calculate the Respective Feature Sets}
$\boldsymbol{Y} \leftarrow \boldsymbol{R} \cdot \boldsymbol{X}$
$\boldsymbol{Y}^{(c)} \leftarrow \boldsymbol{R} \cdot \boldsymbol{X}^{(c)}$, $\forall c$

{Learning}
Calculate the derivative $\partial I_2(\boldsymbol{Y}; C)/\partial \boldsymbol{R}$ using Eq. (B.17) and Eq. (B.18)
Update $\boldsymbol{R}$ using Eq. (B.9)

end loop

List of Publications by the Author

Gläser, C. (2011). A computational account on fast and slow mapping during word learning. In *Proceedings of the International Congress for the Study of Child Language*.

Gläser, C., Heckmann, M., Joublin, F., and Goerick, C. (2008a). Auditory-based formant estimation in noise using a probabilistic framework. In *Proceedings of the Interspeech*, pages 2606–2609.

Gläser, C., Heckmann, M., Joublin, F., and Goerick, C. (2010a). Combining auditory preprocessing and Bayesian estimation for robust formant tracking. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):224–236.

Gläser, C., Heckmann, M., Joublin, F., and Goerick, C. (2010b). Robust formant tracking in echoic noisy environments. In *Proceedings of the ITG Conference on Speech Communication*.

Gläser, C., Heckmann, M., Joublin, F., Goerick, C., and Gross, H.-M. (2007). Joint estimation of formant trajectories via spectro-temporal smoothing and Bayesian techniques. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages IV–477–480.

Gläser, C. and Joublin, F. (2010a). An adaptive normalized Gaussian network and its application to online category learning. In *Proceedings of the International Joint Conference on Neural Networks*, pages 675–682.

Gläser, C. and Joublin, F. (2010b). A computational model for grounding words in the perception of agents. In *Proceedings of the International Conference on Development and Learning*, pages 26–32. Best Paper Award.

Gläser, C. and Joublin, F. (2010c). Perceptually grounded word meaning acquisition: A computational model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1744–1749.

Gläser, C. and Joublin, F. (in press). Firing rate homeostasis for dynamic neural field formation. *IEEE Transactions on Autonomous Mental Development*.

Gläser, C., Joublin, F., and Goerick, C. (2008b). Enhancing topology preservation during neural field development via wiring length minimization. In Kurkova, V., Neruda, R., and Koutnik, J., editors, *Artificial Neural Networks - ICANN 2008, Part I*, volume 5163 of *Lecture Notes in Computer Science*, pages 593–602. Springer.

Gläser, C., Joublin, F., and Goerick, C. (2008c). Homeostatic development of dynamic neural fields. In *Proceedings of the International Conference on Development and Learning*, pages 121–126.

Gläser, C., Joublin, F., and Goerick, C. (2009a). Intrinsically regulated self-organization of topologically ordered neural maps. In *Frontiers in Computational Neuroscience. Bernstein Conference on Computational Neuroscience.*

Gläser, C., Joublin, F., and Goerick, C. (2009b). Learning and use of sensorimotor schemata maps. In *Proceedings of the International Conference on Development and Learning*, pages 1–8.

Heckmann, M. and Gläser, C. (2011). Discriminant sub-space projection of spectro-temporal speech features based on maximizing mutual information. In *Proceedings of the Interspeech.*

Heckmann, M., Gläser, C., Joublin, F., and Nakadai, K. (2010a). Applying geometric source separation for improved pitch extraction in human-robot interaction. In *Proceedings of the Interspeech*, pages 34–39.

Heckmann, M., Gläser, C., Joublin, F., Yamamoto, S., and Nakadai, K. (2010b). Pitch extraction for interaction with ASIMO. In *Proceedings of the HRI Global Workshop*, pages 34–39.

Heckmann, M., Gläser, C., Vaz, M., Rodemann, T., Joublin, F., and Goerick, C. (2008). Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 1699–1704.

Rodemann, T., Heckmann, M., Gläser, C., Joublin, F., and Goerick, C. (2010). Towards speech acquisition in natural interaction on ASIMO. *Journal of the Robotics Society of Japan*, 28(1):18–22.

Bibliography

Alberini, C. M. (2011). The role of reconsolidation and the dynamic process of long-term memory formation and storage. *Frontiers in Behavioral Neuroscience*, 5(12):1–10.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87.

Andersen, R. A. and Buneo, C. A. (2002). Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, 25:189–220.

Au, T. K. and Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child Development*, 61(5):1474–1490.

Bahrick, L. E. and Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, 21(6):963–973.

Ballard, D. H. (1997). *An Introduction to Natural Computation*. MIT Press.

Baseler, H. A., Morland, A. B., and Wandell, B. A. (1999). Topographic organization of human visual areas in the absence of input from primary cortex. *The Journal of Neuroscience*, 19(7):2619–2627.

Bayley, P. J. and Squire, L. R. (2005). Failure to acquire new semantic knowledge in patients with large medial temporal lobe lesions. *Hippocampus*, 15(2):273–280.

Berlin, B. and Kay, P. (1991). *Basic Color Terms: Their Universality and Evolution*. University of California Press.

Berridge, M. J. (1998). Neuronal calcium signaling. *Neuron*, 21(1):13–26.

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., and Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience*, 17(1):353–362.

Blakemore, C. and Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research*, 15(4):439–440.

Bliss, T. V. and Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39.

Bloom, P. (2000). *How Children Learn the Meaning of Words*. MIT Press.

Bloom, P. and Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences*, 2(2):67–73.

Bohannon, J. N. and Stanowicz, L. B. (1988). The issue of negative evidence: Adult responses to children's language errors. *Developmental Psychology*, 24(5):684–689.

Bonhoeffer, T. and Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature*, 353(6343):429–431.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, 25:151–188.

Booth, A. E. (2006). Object function and categorization in infancy: Two mechanisms of facilitation. *Infancy*, 10(2):145–169.

Booth, A. E. (2009). Causal supports for early word learning. *Child Development*, 80(4):1243–1250.

Booth, A. E. and Waxman, S. (2002a). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, 38(6):948–957.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In Hawkins, J. A., editor, *Explaining Language Universals*, pages 73–101. Wiley-Blackwell.

Bowerman, M. and Choi, S. (2003). Space under construction: Language-specific spatial categorization in first language acquisition. In Gentner, D. and Goldin-Meadow, S., editors, *Language in Mind: Advances in the Study of Language and Thought*, pages 387–427. MIT Press.

Breitenstein, C., Jansen, A., Deppe, M., Foerster, A.-F., Sommer, J., Wolbers, T., and Knecht, S. (2005). Hippocampus activity differentiates good from poor learners of a novel lexicon. *Neuroimage*, 25(3):958–968.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., and Pylkkänen, L. (in press). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*.

Broca, P. (1861). Remarque sur le siege de la faculté du language articulé, suivie d'une observation d'aphémie (perte de la parole). *Bulletin de la société anatomique de Paris*, 36:330–356.

Carey, S. (1978). The child as word learner. In Halle, M., Brsnan, J., and Miller, A., editors, *Linguistic Theory and Psychological Reality*, pages 264–293. MIT Press.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, 6(3):184–205.

Carey, S. and Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29.

Carr, M. F., Jadhav, S. P., and Frank, L. M. (2011). Hippocampal replay in the awake state: A potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, 14(2):147–153.

Catani, M., Allin, M. P. G., Husain, M., Pugliese, L., Mesulam, M. M., Murray, R. M., and Jones, D. K. (2007). Symmetries in human brain language pathways correlate with verbal recall. *PNAS*, 104(43):17163–17168.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: A library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm (14.06.2011).

Chen, B. L., Hall, D. H., and Chklovskii, D. B. (2006). Wiring optimization can relate neuronal structure and function. *PNAS*, 103(12):4723–4728.

Cherniak, C., Mokhtarzada, Z., Rodriguez-Esteban, R., and Changizi, K. (2004). Global optimization of cerebral cortex layout. *PNAS*, 101(4):1081–1086.

Chklovskii, D. B. (2004). Synaptic connectivity and neuronal morphology: Two sides of the same coin. *Neuron*, 43(5):609–617.

Chouinard, M. M. and Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–669.

Clay, F., Bowers, J. S., Davis, C. J., and Hanley, D. A. (2007). Teaching adults new words: The role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):970–976.

Cohen, Y. E. and Andersen, R. A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience*, 3(7):553–562.

Coombes, S. (2005). Waves, bumps, and patterns in neural field theories. *Biological Cybernetics*, 93(2):91–108.

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., and Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, 92(1-2):179–229.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, 5(9):382–387.

Davies, I. R., Ozgen, E., Pilling, M., and Wiggett, A. (2003). Categorical perception, perceptual magnet and prototype-bias: Same or different phenomena? *Journal of Vision*, 3(9):250.

Davis, J. V. and Dhillon, I. (2006). Differential entropic clustering of multivariate Gaussians. In *Advances in Neural Information Processing Systems 19*. MIT Press.

Davis, M. H. and Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3773–3800.

Dayan, P. (1999). Unsupervised learning. In Wilson, R. A. and Keil, F. C., editors, *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press.

Deng, L., Cui, X., Pruvenok, R., Chen, Y., Momen, S., and Alwan, A. (2006). A database of vocal tract resonance trajectories for research in speech processing. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages I – 369–372.

Desai, N. S. (2003). Homeostatic plasticity in the CNS: synaptic and intrinsic forms. *Journal of Physiology – Paris*, 97(4-6):391–402.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999a). BDNF regulates the intrinsic excitability of cortical neurons. *Learning & Memory*, 6(3):284–291.

Desai, N. S., Rutherford, L. C., and Turrigiano, G. G. (1999b). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience*, 2(6):515–520.

DeSieno, D. (1988). Adding a conscience to competitive learning. In *Proceedings of the International Conference on Neural Networks*, pages 117–124.

Dewar, K. and Xu, F. (2007). Do 9-month-old infants expect distinct words to refer to kinds? *Developmental Psychology*, 43(5):1227–1238.

Dickinson, R. K. (1988). Learning names for materials: Factors constraining and limiting hypotheses about word meaning. *Cognitive Development*, 3(1):15–35.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. John Wiley & Sons.

Erdogmus, D., Hild, K., and Principe, J. (2003). Online entropy manipulation: Stochastic information gradient. *IEEE Signal Processing Letters*, 10(8):242–245.

Feldman, J. and Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain and Language*, 89(2):385–392.

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*, 81(2):472–479.

Foeller, E. and Feldman, D. E. (2004). Synaptic basis for developmental plasticity in somatosensory cortex. *Current Opinion in Neurobiology*, 14(1):89–95.

Fogel, D. B. (1994). An introduction to simulated evolutionary optimization. *IEEE Transactions on Neural Networks*, 5(1):3–14.

Fontanari, J. F., Tikhanoff, V., Cangelosi, A., Ilin, R., and Perlovsky, L. I. (2009). Cross-situational learning of object-word mapping using neural modeling fields. *Neural Networks*, 22(5-6):579–585.

Frankland, P. W. and Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, 6(2):119–130.

Frey, P. W. and Slate, D. J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, 6:161–182.

Friedrich, M. and Friederici, A. D. (2008). Neurophysiological correlates of online word learning in 14-month-old infants. *Neuroreport*, 19(18):1757–1761.

Friedrich, M. and Friederici, A. D. (in press). Word learning in 6-month-olds: Fast encoding - weak retention. *Journal of Cognitive Neuroscience*.

Fulkerson, A. L., Waxman, S. R., and Seymour, J. M. (2006). Linking object names and object categories: Words (but not tones) facilitate object categorization in 6- and 12-month-olds. In *Proceedings of the 30th Annual Boston University Conference on Language Development*.

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, 80(4):1016–1025.

Ganger, J. and Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, 40(4):621–632.

Gentner, D. (2006). Why verbs are hard to learn. In Hirsh-Pasek, K. and Golinkoff, R., editors, *Action Meets Word: How Children Learn Verbs*, pages 544–564. Oxford University Press.

Gershkoff-Stowe, L. and Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*, 50(3):682–697.

Ghahramani, Z. (2004). Unsupervised learning. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*. Springer.

Gleitman, L. R. and Newport, E. L. (1995). The invention of language by children: Environmental and biological influences on the acquisition of language. In Gleitman, L. R. and Liberman, M., editors, *An Invitation to Cognitive Science: Language*. MIT Press.

Gläser, C. (2011). A computational account on fast and slow mapping during word learning. In *Proceedings of the International Congress for the Study of Child Language*.

Gläser, C., Heckmann, M., Joublin, F., and Goerick, C. (2010a). Combining auditory preprocessing and Bayesian estimation for robust formant tracking. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):224–236.

Gläser, C. and Joublin, F. (2010a). An adaptive normalized Gaussian network and its application to online category learning. In *Proceedings of the International Joint Conference on Neural Networks*, pages 675–682.

Gläser, C. and Joublin, F. (2010b). A computational model for grounding words in the perception of agents. In *Proceedings of the International Conference on Development and Learning*, pages 26–32.

Gläser, C. and Joublin, F. (2010c). Perceptually grounded word meaning acquisition: A computational model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1744–1749.

Gläser, C. and Joublin, F. (in press). Firing rate homeostasis for dynamic neural field formation. *IEEE Transactions on Autonomous Mental Development*.

Gläser, C., Joublin, F., and Goerick, C. (2008b). Enhancing topology preservation during neural field development via wiring length minimization. In Kurkova, V., Neruda, R., and Koutnik, J., editors, *Artificial Neural Networks - ICANN 2008, Part I*, volume 5163 of *Lecture Notes in Computer Science*, pages 593–602. Springer.

Gläser, C., Joublin, F., and Goerick, C. (2008c). Homeostatic development of dynamic neural fields. In *Proceedings of the International Conference on Development and Learning*, pages 121–126.

Gläser, C., Joublin, F., and Goerick, C. (2009a). Intrinsically regulated self-organization of topologically ordered neural maps. In *Frontiers in Computational Neuroscience. Bernstein Conference on Computational Neuroscience.*

Gómez, R. L., Bootzin, R. R., and Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science*, 17(8):670–674.

Goerick, C., Schmuedderich, J., Bolder, B., Janssen, H., Gienger, M., Bendig, A., Heckmann, M., Rodemann, T., Brandl, H., Domont, X., and Mikhailova, I. (2009). Interactive online multimodal association for internal concept building in humanoids. In *Proceedings of the International Conference on Humanoids*, pages 411–418.

Gold, K., Doniec, M., Crick, C., and Scassellati, B. (2009). Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence*, 173:145–166.

Golinkoff, R. M., Mervis, C. B., and Hirsh-Pasek, K. (1994). Early object labels: the case for a developmental lexical principles framework. *Journal of Child Language*, 21(1):125–155.

Halligan, P. W., Marshall, J. C., Wade, D. T., Davey, J., and Morrison, D. (1993). Thumb in cheek? Sensory reorganization and perceptual plasticity after limb amputation. *Neuroreport*, 4(3):233–236.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 32(1–3):335–346.

Hart, B. and Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, 27(1):4–9.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

Heckmann, M., Gläser, C., Vaz, M., Rodemann, T., Joublin, F., and Goerick, C. (2008). Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 1699–1704.

Herrmann, M. (1995). Self-organizing feature maps with self-organizing neighborhood widths. In *Proceedings of the International Conference on Neural Networks*, pages 2998–3003.

Hickok, G. and Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Hild, K., Erdogmus, D., Torkkola, K., and Principe, J. (2006). Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1385–1392.

Hoffman, A. B., Harris, H. D., and Murphy, G. L. (2008). Prior knowledge enhances the category dimensionality effect. *Memory & Cognition*, 36(2):256–270.

Holland, R. and Ralph, M. A. L. (2010). The anterior temporal lobe semantic hub is a part of the language neural network: Selective disruption of irregular past tense verbs by rTMS. *Cerebral Cortex*, 20(12):2771–2775.

Hollich, G., Golinkoff, R. M., and Hirsh-Pasek, K. (2007). Young children associate novel words with complex objects rather than salient parts. *Developmental Psychology*, 43(5):1051–1061.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., and Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 65(3):i–vi, 1–123.

Hooser, S. D. V., Heimel, J. A. F., Chung, S., Nelson, S. B., and Toth, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *The Journal of Neuroscience*, 25(1):19–28.

Horst, J. S., Oakes, L. M., and Madole, K. L. (2005). What does it look like and what can it do? Category structure influences how infants categorize. *Child Development*, 76(3):614–631.

Horst, J. S. and Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, 13(2):128–157.

Huang, G.-B., Saratchandran, P., and Sundararajan, N. (2005). A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation. *IEEE Transactions on Neural Networks*, 16(1):57–67.

Imai, M., Gentner, D., and Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, 9(1):45–75.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Ji, D. and Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, 10(1):100–107.

Joublin, F., Spengler, F., Wacquant, S., and Dinse, H. R. (1996). A columnar model of somatosensory reorganizational plasticity based on Hebbian and non-Hebbian learning rules. *Biological Cybernetics*, 74(3):275–286.

Kaas, J. H., Nelson, R. J., Sur, M., Lin, C. S., and Merzenich, M. M. (1979). Multiple representations of the body within the primary somatosensory cortex of primates. *Science*, 204(4392):521–523.

Kadirkamanathan, V. and Niranjan, M. (1993). A function estimation approach to sequential learning with neural networks. *Neural Computation*, 5(6):954–975.

Kaminski, J., Call, J., and Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping". *Science*, 304(5677):1682–1683.

Kaplan, F. (1998). A new approach to class formation in multi-agent simulations of language evolution. In *Proceedings of the International Conference on Multi Agent Systems*, pages 158–165.

Kay, P. and Regier, T. (2006). Language, thought and color: Recent developments. *Trends in Cognitive Sciences*, 10(2):51–54.

Kemmerer, D., Castillo, J. G., Talavage, T., Patterson, S., and Wiley, C. (2008). Neuroanatomical distribution of five semantic components of verbs: Evidence from fMRI. *Brain and Language*, 107(1):16–43.

Kindermann, R. and Snell, J. L. (1980). Markov random fields and their applications. In *Contemporary Mathematics*. American Mathematical Society.

Kirk, J. and Zurada, J. (2000). A two-stage algorithm for improved topography preservation in self-organizing maps. In *Proceedings of the International Conference on Systems, Man, and Cybernetics*, volume 4, pages 2527–2532.

Kirstein, S., Wersing, H., Gross, H. M., and Körner, E. (2009). An integrated system for incremental learning of multiple visual categories. In *Advances in Neuro-Information Processing*, volume 5506 of *Lecture Notes in Computer Science*, pages 813–820. Springer.

Kiviluoto, K. (1996). Topology preservation in self-organizing maps. In *Proceedings of the International Conference on Neural Networks*, volume 1, pages 294–299.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1988). The neural phonetic typewriter. *Computer*, 21(3):11–22.

Kokaia, Z., Bengzon, J., Metsis, M., Kokaia, M., Persson, H., and Lindvall, O. (1993). Coexpression of neurotrophins and their receptors in neurons of the central nervous system. *PNAS*, 90(14):6711–6715.

Kopp, S., Gesellensetter, L., Krämer, N., and Wachsmuth, I. (2005). A conversational agent as museum guide: Design and evaluation of a real-world application. In *Intelligent Virtual Agents*, volume 3661 of *Lecture Notes in Computer Science*, pages 329–343. Springer.

Kovack-Lesh, K. A., Horst, J. S., and Oakes, L. M. (2008). The cat is out of the bag: The joint influence of previous experience and looking behavior on infant categorization. *Infancy*, 13(4):285–307.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44.

Law, J. (2009). *Modeling the Development of Organization for Orientation Preference in Primary Visual Cortex*. PhD thesis, University of Edinburgh.

Lemon, R. (1988). The output map of the primate motor cortex. *Trends in Neurosciences*, 11(11):501–506.

Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision*. Springer-Verlag.

Lindsay, Shane; Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning*, 60:45–63.

Lledo, P.-M., Alonso, M., and Grubb, M. S. (2006). Adult neurogenesis and functional plasticity in neuronal circuits. *Nature Reviews Neuroscience*, 7(3):179–193.

Lu, Y., Sundararajan, N., and Saratchandran, P. (1997). A sequential learning scheme for function approximation using minimal radial basis function neural networks. *Neural Computation*, 9(2):461–478.

Lyn, H. and Savage-Rumbaugh, E. S. (2000). Observational word learning in two bonobos (pan paniscus): Ostensive and non-ostensive contexts. *Language & Communication*, 20(3):255–273.

Mandler, J. M. (2004). *The Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

Marder, E. and Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563–574.

Marder, E. and Prinz, A. A. (2002). Modeling stability in neuron and network function: The role of activity in homeostasis. *Bioessays*, 24(12):1145–1154.

Markman, E. (1994). Constraints on word meaning in early language acquisition. *Lingua*, 92(1-4):199–227.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science: A Multidisciplinary Journal*, 14(1):57–77.

Markman, E. M. and Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2):121–157.

Markson, L. and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385(6619):813–815.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45.

Mayor, J. and Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, 117(1):1–31.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

McDonald, C. T. and Burkhalter, A. (1993). Organization of long-range inhibitory connections with rat visual cortex. *The Journal of Neuroscience*, 13(2):768–781.

McDonough, C., Song, L., Pasek, K. H., Golinkoff, R. M., and Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, 14(2):181–189.

McDonough, L., Choi, S., and Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, 46(3):229–259.

Mervis, C. (1987). Child-basic object categories and early lexical development. In Neisser, U., editor, *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, pages 201–233. Cambridge University Press.

Mervis, C., Catlin, J., and Rosch, E. (1975). Development of the structure of color categories. *Developmental Psychology*, 11(1):54–60.

Mestres-Missé, A., Càmara, E., Rodriguez-Fornells, A., Rotte, M., and Münte, T. F. (2008). Functional neuroanatomy of meaning acquisition from context. *Journal of Cognitive Neuroscience*, 20(12):2153–2166.

Mestres-Missé, A., Rodriguez-Fornells, A., and Münte, T. F. (2010). Neural differences in the mapping of verb and noun concepts onto novel words. *Neuroimage*, 49(3):2826–2835.

Mikhailova, I. and Goerick, C. (2005). Conditions of activity bubble uniqueness in dynamic neural fields. *Biological Cybernetics*, 92(2):82–91.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294.

Morgan, R. and Rochat, P. (1997). Intermodal calibration of the body in early infancy. *Ecological Psychology*, 9(1):1–23.

Myrvoll, T. A. and Soong, F. K. (2003). On divergence based clustering of normal distributions and its application to HMM adaptation. In *Proceedings of the Eurospeech*, pages 1517–1520.

Nation, P. (1993). Vocabulary size, growth, and use. In Schreuder, R. and Weltens, B., editors, *The Bilingual Lexicon*, pages 115–134. John Benjamins Publishing Company.

Nelson, D. G. K., Russell, R., Duke, N., and Jones, K. (2000). Two-year-olds will name artifacts by their functions. *Child Development*, 71(5):1271–1288.

Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review*, 81(4):267–285.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions. *Bulletin of the Japanese Society for Fish Science*, 22:526–530.

Ohki, K. and Reid, R. C. (2007). Specificity and randomness in the visual cortex. *Current Opinion in Neurobiology*, 17(4):401–407.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273.

O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455–462.

O'Reilly, R. and Norman, K. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, 6(12):505–510.

O'Reilly, R. and Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, 10(4):389–397.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987.

Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. In *Proceedings of the Symposium on Hearing, Auditory Physiology and Perception*, pages 429–446.

Perone, S., Madole, K. L., Ross-Sheehy, S., Carey, M., and Oakes, L. M. (2008). The relation between infants' activity with objects and attention to object appearance. *Developmental Psychology*, 44(5):1242–1248.

Phaf, R., Dulk, P. D., Tijsseling, A., and Lebert, E. (2001). Novelty-dependent learning and topological mapping. *Connection Science*, 13(4):293–321.

Platt, J. (1991). A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225.

Plunkett, K. (1997). Theories of early language acquisition. *Trends in Cognitive Sciences*, 1(4):146–153.

Plunkett, K., Hu, J.-F., and Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2):665–681.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *Journal of Anatomy*, 197(3):335–359.

Pulvermüller, F., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P. J., and Patterson, K. (2010). The word processing deficit in semantic dementia: All categories are equal, but some categories are more equal than others. *Journal of Cognitive Neuroscience*, 22(9):2027–2041.

Pulvermüller, F. and Hauk, O. (2006). Category-specific conceptual processing of color and form in left fronto-temporal cortex. *Cerebral Cortex*, 16(8):1193–1201.

Pulvermüller, F., Härle, M., and Hummel, F. (2000). Neurophysiological distinction of verb categories. *Neuroreport*, 11(12):2789–2793.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427.

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Quinn, P. C. (2002). Category representation in young infants. *Current Directions in Psychological Science*, 11(2):66–70.

Regier, T. (1990). Learning spatial terms without explicit negative evidence. Technical Report TR-90-057, International Computer Science Institute, Berkeley, CA.

Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. MIT Press.

Regier, T. (2005). The emergence of words : Attentional learning in form and meaning. *Cognitive Science: A Multidisciplinary Journal*, 29(6):819–865.

Regier, T. and Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, 93(2):147–155.

Renyi, A. (1970). *Probability Theory*. North-Holland Publishing Company.

Richmond, J. and Nelson, C. A. (2007). Accounting for change in declarative memory: A cognitive neuroscience perspective. *Developmental Review*, 27(3):349–373.

Roberson, D., Davidoff, J., Davies, I., and Shapiro, L. R. (2004). The development of color categories in two languages: a longitudinal study. *Journal of Experimental Psychology: General*, 133(4):554–571.

Roberson, D., Davies, I., and Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3):369–398.

Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., and de Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3711–3735.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science: A Multidisciplinary Journal*, 26(1):113–146.

Rutherford, L. C., DeWan, A., Lauer, H. M., and Turrigiano, G. G. (1997). Brain-derived neurotrophic factor mediates the activity-dependent regulation of inhibition in neocortical cultures. *The Journal of Neuroscience*, 17(12):4527–4535.

Rutherford, L. C., Nelson, S. B., and Turrigiano, G. G. (1998). BDNF has opposite effects on the quantal amplitude of pyramidal neuron and interneuron excitatory synapses. *Neuron*, 21(3):521–530.

Samejima, K. and Omori, T. (1999). Adaptive internal state space construction method for reinforcement learning of a real-world agent. *Neural Networks*, 12(7-8):1143–1155.

Santhakumar, V. and Soltesz, I. (2004). Plasticity of interneuronal species diversity and parameter variance in neurological diseases. *Trends in Neurosciences*, 27(8):504–510.

Sato, M. and Ishii, S. (2000). On-line EM algorithm for the Normalized Gaussian Network. *Neural Computation*, 12(2):407–432.

Schreiner, C. E. (1992). Functional organization of the auditory cortex: Maps and mechanisms. *Current Opinion in Neurobiology*, 2(4):516–521.

Schyns, P. G. and Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3):681–696.

Sekino, M., Katagami, D., and Nitta, K. (2005). State space self-organization based on the interaction between basis functions. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 2929–2934.

Sirosh, J. and Miikkulainen, R. (1994). Cooperative self-organization of afferent and lateral connections in cortical maps. *Biological Cybernetics*, 71:65–78.

Smith, L. B. (1995). Self-organizing processes in learning to learn words: Development is not induction. In Nelson, C. A., editor, *Basic and Applied Perspectives on Learning, Cognition, and Development*. Lawrence Erlbaum Associates.

Smith, L. B., Jones, S. S., and Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, 60(2):143–171.

Smith, L. B. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, 99(2):195–231.

Steels, L. and Kaplan, F. (2002). Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32.

Sullivan, T. J. and de Sa, V. R. (2006). Homeostatic synaptic scaling in self-organizing maps. *Neural Networks*, 19(6-7):734–743.

Takashima, A., Petersson, K. M., Rutters, F., Tendolkar, I., Jensen, O., Zwarts, M. J., McNaughton, B. L., and Fernández, G. (2006). Declarative memory consolidation in humans: A prospective functional magnetic resonance imaging study. *PNAS*, 103(3):756–761.

Tarullo, A. R., Balsam, P. D., and Fifer, W. P. (2011). Sleep and infant learning. *Infant and Child Development*, 20(1):35–46.

Taylor, J. G. (1999). Neural 'bubble' dynamics in two dimensions: foundations. *Biological Cybernetics*, 80:393–409.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63(2):411–423.

Tkalcic, M. and Tasic, J. (2003). Colour spaces: Perceptual, historical and applicational background. In *Proceedings of the Eurocon, Computer as a Tool*, pages 304–308.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Torkkola, K. (2003). Feature extraction by non parametric mutual information maximization. *The Journal of Machine Learning Research*, 3(7–8):1415–1438.

Triesch, J. (2007). Synergies between intrinsic and synaptic plasticity in individual model neurons. *Neural Computation*, 19(4):885–909.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–896.

Turrigiano, G. G. and Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology*, 10(3):358–364.

Turrigiano, G. G. and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2):97–107.

Vaz, M., Brandl, H., Joublin, F., and Goerick, C. (2009). Speech imitation with a child's voice: Addressing the correspondence problem. In *Proceedings of the International Conference on Speech and Computer*.

Verfaellie, M., Koseff, P., and Alexander, M. P. (2000). Acquisition of novel semantic information in amnesia: Effects of lesion location. *Neuropsychologia*, 38(4):484–492.

Vijayakumar, S., D'Souza, A., and Schaal, S. (2005). Incremental online learning in high dimensions. *Neural Computation*, 17(12):2602–2634.

Villmann, T., Der, R., Herrmann, M., and Martinetz, T. M. (1997). Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266.

Wang, F., Nemes, A., Mendelsohn, M., and Axel, R. (1998). Odorant receptors govern the formation of a precise topographic map. *Cell*, 93(1):47–60.

Wang, S.-H. and Morris, R. G. M. (2010). Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annual Review of Psychology*, 61:49–79, C1–4.

Ware, E. A. and Booth, A. E. (2010). Form follows function: Learning about function helps children learn about shape. *Cognitive Development*, 25(2):124–137.

Warrens, M. J. (2008). *Similarity Coefficients for Binary Data: Properties of Coefficients, Coefficient Matrices, Multi-way Metrics and Multivariate Coefficients*. PhD thesis, Leiden University.

Waxman, S. R. and Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3):257–302.

Weizman, Z. O. and Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, 37(2):265–279.

Wellens, P., Loetzsch, M., and Steels, L. (2008). Flexible word meaning in embodied agents. *Connection Science*, 20(2–3):173–191.

Wernicke, C. (1874). *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Wilhelm, I., Diekelmann, S., and Born, J. (2008). Sleep in children improves memory performance on declarative but not procedural tasks. *Learning & Memory*, 15(5):373–377.

Wilhelm, J. C., Rich, M. M., and Wenner, P. (2009). Compensatory changes in cellular excitability, not synaptic scaling, contribute to homeostatic recovery of embryonic network activity. *PNAS*, 106(16):6760–6765.

Wilson, H. R. and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13(2):55–80.

Wu, R., Mareschal, D., and Rakison, D. H. (2011). Attention to multiple cues during spontaneous object labeling. *Infancy*, 16(5):545–556.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.

Xu, L. (1998). RBF nets, mixture experts, and Bayesian Ying-Yang learning. *Neurocomputing*, 19:223–257.

Yeung, H. H. and Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113(2):234–243.

Young, S. J. and Woodland, P. C. (1993). The use of state tying in continuous speech recognition. In *Proceedings of the Eurospeech*, pages 2203–2206.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

Zeki, S., Watson, J. D., Lueck, C. J., Friston, K. J., Kennard, C., and Frackowiak, R. S. (1991). A direct demonstration of functional specialization in human visual cortex. *The Journal of Neuroscience*, 11(3):641–649.

Zhang, W. and Linden, D. J. (2003). The other side of the engram: Experience-driven changes in neuronal intrinsic excitability. *Nature Reviews Neuroscience*, 4(11):885–900.