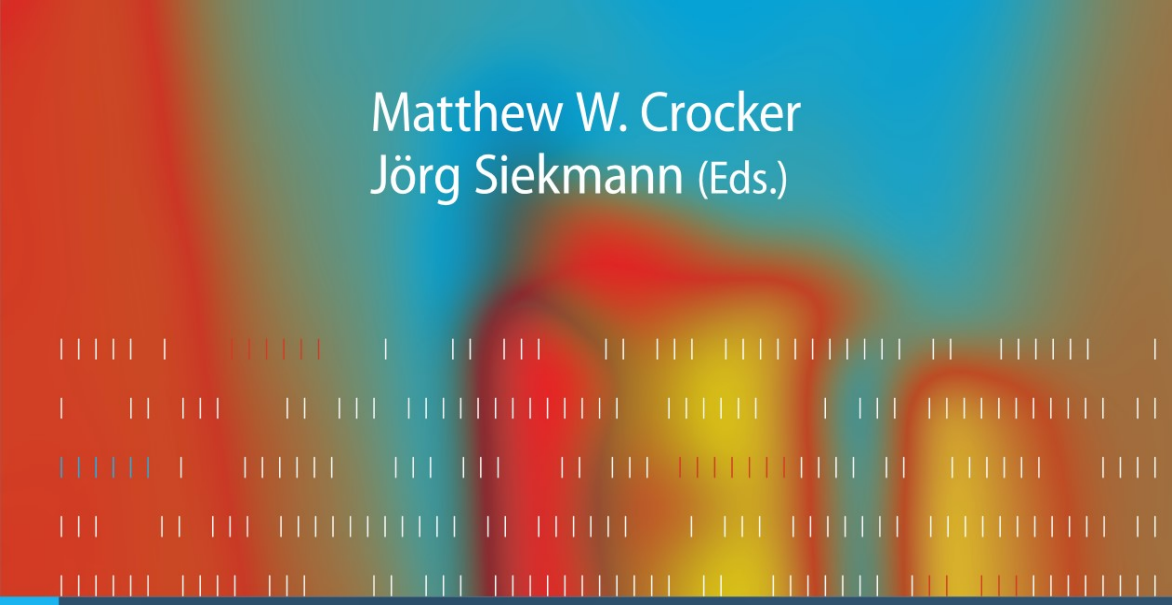
Matthew W. Crocker
Jörg Siekmann (Eds.)



Resource-Adaptive Cognitive Processes

Springer

Cognitive Technologies

Matthew W. Crocker · Jörg Siekmann (Eds.)

Resource-Adaptive Cognitive Processes

With 148 Figures and 14 Tables

Springer

Editors
Prof. Dr. Matthew W. Crocker
Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
crocker@coli.uni-sb.de

Prof. Dr. Jörg Siekmann
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
Saarbrücken, Germany
and
Dept. of Computer Science
Saarland University
Saarbrücken, Germany
joerg.siekmann@dfki.de

Managing Editors
Prof. Dov M. Gabbay
Augustus De Morgan Professor of Logic
Department of Computer Science
King's College London
Strand, London WC2R 2LS, UK
dov.gabbay@kcl.ac.uk

Prof. Dr. Jörg Siekmann
Deutsches Forschungszentrum
für Künstliche Intelligenz (DFKI)
Saarbrücken, Germany
and
Dept. of Computer Science
Saarland University
Saarbrücken, Germany
joerg.siekmann@dfki.de

Cognitive Technologies ISSN 1611-2482
ISBN 978-3-540-89407-0 e-ISBN 978-3-540-89408-7
DOI 10.1007/978-3-540-89408-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009938631

ACM Computing Classification (1998): I.2, H.1, H.5, J.4

Cover design: KünkelLopka GmbH, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The contributions to this volume are drawn from the interdisciplinary research carried out within the *Sonderforschungsbereich* (SFB 378), a special long-term funding scheme of the German National Science Foundation (DFG). *Sonderforschungsbereich 378* was situated at Saarland University, with colleagues from artificial intelligence, computational linguistics, computer science, philosophy, psychology – and in its final phases – cognitive neuroscience and psycholinguistics.

The funding covered a period of 12 years, which was split into four phases of 3 years each, ending in December of 2007. Every sub-period culminated in an intensive reviewing process, comprising written reports as well as on-site presentations and demonstrations to the external reviewers. We are most grateful to these reviewers for their extensive support and critical feedback; they contributed their time and labor freely to the DFG,[1] the independent and self-organized institution of German scientists. The final evaluation of the DFG reviewers judged the overall performance and the actual work with the highest possible mark, i.e. "excellent".

All contributions were written especially for this volume and summarize 12 years of research that has resulted in several hundred individual publications. They represent in our opinion the most important subset of the many individual projects and offer an overarching perspective not reflected in the individual scientific publications. Specifically, contributors sought to present their results in a summary fashion covering the main findings of this long period, while also making clear the technical and scientific contribution to the overarching theme of the volume: "resource-adaptive cognitive processes". Each paper was reviewed by an internal and external expert in the specific subject area, and finally by the two editors. We are indebted to these reviewers, who offered their time to review each of the contributions.

Finally, we are also most grateful for the intense collaboration with our colleagues over this long time span of more than a decade: Hans Engelkamp (Psychology), Ulman Lindenberger (Psychology), Kuno Lorenz (Philosophy), Axel Mecklinger (Neuroscience), Manfred Pinkal (Computational Linguistics), Gert Smolka

[1] This translates literally into "German Community of Scientists"

(Computer Science), Werner Tack (Psychology), Hans Uszkoreit (Computational Linguistics), Wolfgang Wahlster (AI), Hubert Zimmer (Psychology). Our own research areas are psycholinguistics (Matthew Crocker) and automated reasoning and AI (Jörg Siekmann).

Saarbrücken, Germany

Matthew W. Crocker
Jörg Siekmann

Contents

Contributors

Serge Autexier DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany, autexier@dfki.de

Jörg Baus Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, baus@cs.uni-sb.de

Christoph Benzmüller Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, chris@ags.uni-sb.de

Mark Buckley Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany, buckley@coli.uni-sb.de

Matthew W. Crocker Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrucken, Germany, crocker@coli.uni-sb.de

Ralph Debusmann Programming Systems Lab, Saarland University, 66123 Saarbrücken, Germany, rade@ps.uni-sb.de

Dominik Dietrich Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, dietrich@ags.uni-sb.de

Ben Eppinger Developmental Psychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany, eppinger@princeton.edu

Michael Feld DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, Michael.Feld@dfki.de

Nicola K. Ferdinand Experimental Neuropsychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany, n.ferdinand@mx.uni-saarland.de

Patrick Gebhard DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrucken, Germany, Patrick.Gebhard@dfki.de

Paolo Ghisletta Faculty of Psychology and Educational Sciences, University of Geneva, Switzerland, paolo.ghisletta@pse.unige.ch

Barbara Großmann-Hutter Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, barbara@cs.uni-sb.de

Dominikus Heckmann DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, heckmann@dfki.de

Helmut Horacek Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, horacek@ags.uni-sb.de

Anthony Jameson DFKI GmbH, Saarbrücken, Germany and Fondazione Bruno Kessler – Istituto per Ricerca Scientifica e Tecnologica (FBK-irst), Trento, Italy, jameson@fbk.eu

Ralf Jung DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, Ralf.Jung@dfki.de

Juergen Kiefer ZMMS – MoDyS Research Group, Technical University Berlin, 10623 Berlin, Germany, juergen.kiefer@zmms.tu-berlin.de

Pia Knoeferle Cognitive Interaction Technology, Bielefeld University, 33615 Bielefeld, Germany, knoeferl@cit-ec.uni-bielefeld.de

Alexander Koller MMCI & Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrucken, Germany, koller@mmci.uni-saarland.de

Jutta Kray Developmental Psychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany, j.kray@mx.uni-saarland.de

Ivana Kruijff-Korbayová Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany, korbay@coli.uni-sb.de

Michael Kruppa DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, Michael.Kruppa@dfki.de

Marco Kuhlmann Department of Linguistics and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden, marco.kuhlmann@lingfil.uu.se

Ulman Lindenberger Center for Lifespan Psychology, Max Planck Institute of Human Development, Berlin, Germany, lindenberger@mpib-berlin.mpg.de

Marshall R. Mayberry School of Social Sciences, Humanities and Arts, UC Merced, CA, USA, marty.mayberry@gmail.com

Axel Mecklinger Experimental Neuropsychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany, mecklinger@mx.uni-saarland.de

Christian Muller Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, cmueller@cs.uni-sb.de

Stefan Münzer Brain and Cognition Unit, Department of Psychology, Saarland University, 66123 Saarbrücken, Germany, s.muenzer@mx.uni-saarland.de

Manfred Pinkal Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany, pinkal@coli.uni-sb.de

Ralf Rummer Department of Psychology, Erfurt University, 99105 Erfurt, Germany, ralf.rummer@unierfurt.de

Marvin Schiller Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, schiller@ags.uni-sb.de

Michael Schmitz DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany

Tim Schwartz DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany, Tim.Schwartz@dfki.de

Jörg Siekmann DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany, siekmann@dfki.de

Lubomira Spassova DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, Luebomira.Spassova@dfki.de

Christoph Stahl DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany, Christoph.Stahl@dfki.de

Stefan Thater Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany, stth@coli.uni-sb.de

Katja Umla-Runge Brain and Cognition Unit, Department of Psychology, Saarland University, 66123 Saarbrücken, Germany, k.umla-runge@mx.uni-saarland.de

Timo von Oertzen Center for Lifespan Psychology, Max Planck Institute of Human Development, Berlin, Germany, vonoertzen@mpib-berlin.mpg.de

Wolfgang Wahlster DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany, wahlster@dfki.de

Rainer Wasinger Smart Services CRC and The School of Information Technologies, Sydney University, NSW, 2006, Australia, wasinger@it.usyd.edu.au

Andrea Weber Max-Planck-Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands, andrea.weber@mpi.nl

Anja Weiten Experimental Neuropsychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany, ancowei@gmx.de

Frank Wittig SAP AG, Neue Bahnhofstr. 21, D-66386 St. Ingbert, Germany, frank.wittig@sap.com

Magdalena Wolska Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany, magda@coli.uni-sb.de

Hubert Zimmer Brain and Cognition Unit, Department of Psychology, Saarland University, 66123 Saarbrücken, Germany, huzimmer@mx.uni-saarland.de

Resource-Adaptive Cognitive Processes

Jörg Siekmann and Matthew W. Crocker

1 Background

Kaplan and Simon [8] define the objective of cognitive science as the study of *intelligence* emphasizing representation and intelligent behavior as the result of *computation*, which can be analyzed independently of the actual stratum within which it is realized (silicon or protoplasm). Within the constructive computer science paradigm, cognitive processes are shown and analyzed with the help of programs and until recently it has been customary to assume the conventional notion of a serial computation, thus understanding cognitive processes as a chronological sequence of operations. While this understanding appears to be correct for many higher cognitive processes, we extended this view to resource-limited, constraint-based, and concurrent computations, with the aim to integrate the varying research traditions within this approach and thus using the concept of *resource-guided concurrent* computation as a general basis for the examination of resource-adaptive cognitive processes. This general approach is further augmented in some chapters by findings in neuroscience and connectionist models of computation.

This general point of view – which is informed both by results from theoretical computer science and modern programming languages in artificial intelligence (AI) and by the empirical observations of cognitive processes in cognitive psychology and neuroscience – has been exploited for psychological and linguistic problems as well as for decision making under resource limitations. Thus, the focus of the work presented in this volume is the resource adaptation of cognitive processes in general. It is about heuristics, procedures, and cognitive mechanisms that are a natural consequence of adaptation to resource limitations in people as well as in artificially constructed (computer) systems. This book deals with the construction and analysis of resource control in cognitive processes and with the question of how resource adaptation can be presented, analyzed, formalized, and modeled.

J. Siekmann (✉)
DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany
e-mail: siekmann@dfki.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_1,

2 Resource-Adaptive Cognitive Processes

The concept of *limited resources* is paramount in cognitive science (some of the well-known books and articles are listed in the references, see [2–5, 13, 15, 16], and more specifically [6–8]) and it has been applied in many domains. Examples include time and space, (human and machine) memory, knowledge and information, as well as stamina, energy, or tools and instruments. When defining such a notion in specific scientific research areas, it is very often overlooked that this terminology does not necessarily correspond to daily language usage nor to any broader academic use. The development of a terminological framework is thus especially important for interdisciplinary research, involving scientific disciplines with long-standing and complex disciplinary research traditions (cf. [10]).

With this in mind, it is appropriate to initially adopt a relatively broad notion and to view a resource in a first step simply as a "tool" or "source" for accomplishing a task. Since cognitive science investigates intelligent, and very often rational, solving of tasks for natural and artificial systems, the actual tools that are being used for an individual task or goal can be identified and named.

Typical types of resources in everyday life are physical things, human abilities and skills, information, energy, or time. The form in which these resources are used is very diverse: While some resources are consumed when they are used (e.g., electrical energy), others can be used again (e.g., the ability to read). Resources can be used with certain constraints (e.g., economical or safety constraints); they can be used in a strategic manner or without reflection and automatically (cf. [14], p. 55). Some resources can be divided into well-defined parts (e.g., time), while others can only be used as a whole (e.g., a bicycle or a telephone number).

The terminological framework used in the works of this volume comprises several core terms, which are illustrated in Fig. 1. The illustration captures the situation of a (natural or artificial) agent that is to complete both tasks A_1 and A_2 within a given time limit. The agent uses the three resources R_1, R_2, and R_3, where R_1 is not relevant for A_2 as indicated by the arrows. Using the resource R_i for the task A_j is denoted as V_{ij}. The use of V_{ij} can sometimes be marked quantitatively and sometimes only qualitatively; sometimes it can be measured precisely and sometimes it can only be estimated. It is also subject to limitations (see the left column of the figure). For example, for the resource $R_1 = \mathit{place}$ there may be a specified amount F_1, which can be used for A_1 as well as an amount $F_2 = \mathit{time}$, which must be distributed to A_1 and A_2.

$R_3 = \mathit{knowledge}$ is also marked on a one-dimensional scale in this example; however, the limitation is different than for R_1 and R_2. As knowledge is not consumed when using it, the limit is just the "amount of knowledge" available, hence we do not speak of the "consumption of a resource", but just of *usage* such that tools with limitations of this kind can also be understood as resources.

Limits can be more complex – i.e., not additive – like for example the resource *time* under the constraint that switching between tasks leads to a loss of time. Furthermore, limits may not always be global quantitative measures. It may be necessary for example to specify not only the total amount of time available, but the individual time intervals within which a task has to be completed (see [6, 7]).

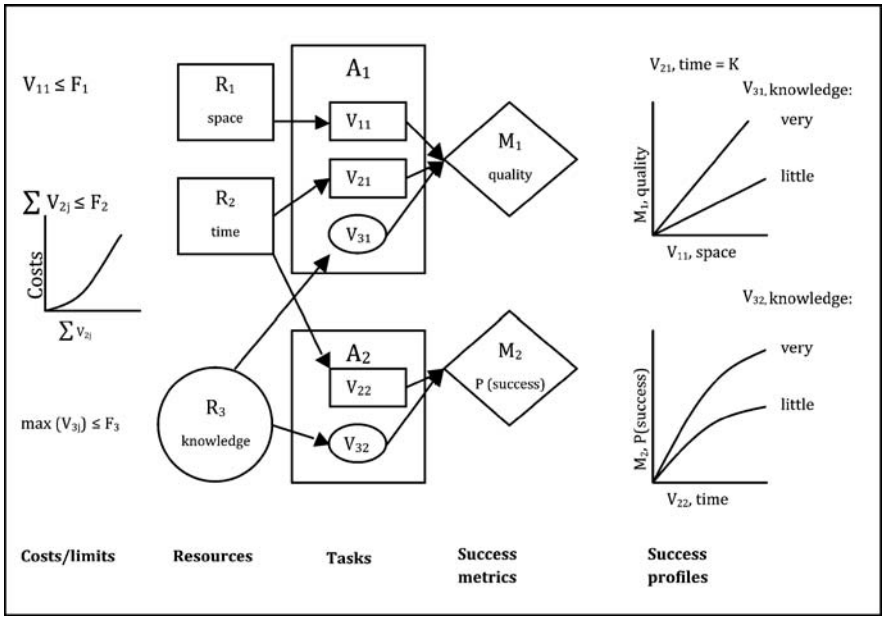


Fig. 1 A terminological framework for resource-adaptive processes

The *costs* of a resource play an important role in the field of business administration, for example. Such a notion makes it possible to express the fact that more use of a resource R_i may be disadvantageous for the agent, even if the limit for R_i is not exceeded. Figure 1 shows on the left hand side the costs of the resource *time*, using a simple cost function.

A central aspect of a resource is that it may not only be restricted but also be useful: The success of an agent carrying out a task A_j must depend positively on the use of the resources for A_j. Various authors (e.g., [1, 11]) have introduced concepts for this purpose, which can be compared to those shown on the right hand side of Fig. 1: *success metrics* and *success profiles*. A *success metric* M_j for a task A_j is a measure for the relative success of the completion of A_j. Sometimes it is necessary to use more than one success metric for a given task, for example, recall and precision or consumption of time versus accuracy. A *success profile* for a task A_j is a function over the use of a resource V_{ij} to the relative success, as shown on the right hand side of Fig. 1. As success often depends on many other – only partly foreseeable – factors, the success profile usually shows only the *expected* success.

Some other definitions in the literature can be seen as special cases of our general framework: Norman and Bobrow [11], for example, consider information not as a resource but rather they differentiate between *process resources* and *data*. For some planning tasks, it may be sensible to consider only those tools as a *resource*, which can be quantified and consumed (see e.g., [12]).

3 What Is a Resource-Adaptive Cognitive Process?

Generally speaking, we call a cognitive process resource-adaptive if the final solution of the task that the process is designed for depends on the available resources and their use. This concept can be further refined into (i) resource-adapting, (ii) resource-adapted, and (iii) resource-limited.

We illustrate these notions in the context of the familiar theory in natural language generation that a speaker intuitively tries to minimize the cognitive load of his working memory (and thus the working memory of the listener).[1]

(i) *Resource-adapting.* The speaker and the listener are in a noisy environment, in a hurry, or otherwise limited in their speech processing capacity. So the speaker will not generate long sentences with say widely distributed relative clauses (see footnote 1) but rather utter a short sentence only. In other words a resource-adapting process takes account of the availability of resources of its own or other agents.

(ii) *Resource-limited.* For example, the speaker tries to offer a complex sentence with some relative clauses, but fails to complete the sentence because of the limits of his working memory. So he just switches to a new and simpler sentence structure. Thus in general, a process is resource-limited if it computes its output without taking any resource limits into account; however, if the computation exceeds the limit it does not simply fail with an error message, but it starts all over again with a simpler technique.

(iii) *Resource-adapted*. This is a process that is fixed; however, it may have undergone some (evolutionary) improvements. In our setting, the speaker simply utters German sentences, but the off-line evaluation of the grammar for everyday spoken German prefers constructions that minimize the cognitive load (of the working memory, the NL-generator, and/or the vocabulary) the speaker may not even be aware of. More generally, an agent acts in a *resource-adapted* way if the cognitive processes developed over a long period reflect cognitive limitations.

Within this very general setting, we can now look at some more specific problems that have been addressed by a number of contributors to this volume.

3.1 What Is a Resource-Limited Process?

Most biological information processing is severely resource-limited as shown in the chapter by Kray and Eppinger, where the more specific topic of age differences is studied. Working memory and visuo-spatial working memory are typical examples as well and are considered in the two chapters by Zimmer and colleagues. Typical techniques for coping with the resource limitations in the area of natural language

[1] See, e.g., H. Uszkoreit et al. [17].

processing are investigated by Koller, Thater, and Pinkal for scope underspecification, and in the chapter by Debusmann and Kuhlmann for parsing with dependency grammars. A rather different simulation technique – Latent Growth Curve Modeling – is critically analyzed by von Oertzen, Ghisletta, and Lindenberger for age-based changes in cognitive resources.

3.2 How Can We Allocate the Given Resources to a Specific Task?

The classical setting is the one first proposed in Doug Lenat's AM system, where each process (in his case each heuristic) obtains a certain amount of computing resources (time and/or space) depending on the overall *heuristic worth* this process is supposed to have [9]. Once this is used up, control is returned to the procedure that invoked this process. A similar control structure is used in the OMEGA system, where Autexier et al. experimented with concurrent computations as well. Another technique is proposed by Jameson et al., where Bayesian nets are used to estimate the user's "resources" (e.g., the amount of time pressure in an airport setting) which in turn determine the system's behavior, in this case for example a way-finding task supported by a navigation system. Related is the chapter by Wittig on learning techniques for Bayesian networks.

3.3 How Can We Allocate Resources That Are Not A Priori Defined?

This is a typical application area for *anytime algorithms*, i.e., very general procedures that work for any amount of resources; however, they may deteriorate when resources become scarce. Jameson et al. provide a good illustration from the area of language generation and understanding processing, where the situated context determines the amount of elaboration. Zimmer, Münzer, and Baus address problems of this kind in augmented cognition, where the navigation system takes the cognitive load (and age limitations) of the user into account (see also Wahlster et al.). Similarly, Benzmüller et al. consider the *cognitive resources* of a mathematics student as a crucial parameter for the performance of the system, and closely related is the chapter by Wolska et al.

3.4 Can We Postulate Indirectly Observable Resources That Help to Explain Experimental Data?

This is particularly important for psychological research, where for example the limitations of the working memory are used to explain the behavior during problem solving tasks. A good entry into this kind of research is provided by Zimmer, Münzer, and Umla-Runge, where a specific visuo-spatial working memory (VSWM)

is postulated to explain the behavioral data, and further neuro-scientific evidence for its actual existence and location in the brain is presented.

3.5 In a Multimodal Context, How Are Diverse Knowledge Resources Exploited?

This is the case where an agent has both knowledge of the world and is also confronted with a specific situation in the world. Weber et al., for example, present experimental findings from human performance in situated natural language understanding, which indicate that comprehension has adapted to prioritize different linking and non-linguistic knowledge as a function of experience over time (frequency of observation) and immediacy (presence in the current situation). Mayberry and Crocker develop a resource-adapted connectionist model that learns not only to prioritize visual context information, but to focus attention on relevant aspects of it. Another interesting though very different case in point is put forward by Ferdinand et al., where error-induced learning is shown as a resource-adaptive process for elderly people, in particular.

4 Structure of This Volume

This volume covers topics from diverse fields such as computer science, psychology, computational linguistics, psycholinguistics, connectionism, neuroscience, as well as artificial intelligence, bound by their common interest in the notion of bounded and adaptive cognitive processes. The contributions are organized into three sections, which we now introduce in turn.

4.1 Part I: Resource-Bounded Cognitive Processes in Human Information Processing

This section of the volume comprises works in psychology, neuroscience, psycholinguistics, computer science, AI, and more generally cognitive science. In *Visuo-spatial Working Memory as a Limited Resource of Cognitive Processing*, Zimmer, Münzer, and Umla-Runge first present an overview of research on human working memory and its limitations, focusing in particular on the visuo-spatial working memory (VSWM) with the claim that this exists distinct from other forms of working memory. Experiments demonstrate how this is used in human problem solving. Further interesting results from neuro-cognitive experiments provide evidence for an anterior–posterior network of inferior occipito-temporal and parietal structures that constitute VSWM. Zimmer, Münzer, and Baus report interdisciplinary work from psychology, computer science, and AI on augmented cognition. *From Resource-Adaptive Navigation Assistance to Augmented Cognition* examines

the interplay of the resource of an assistant system and the resources of the user, by using way-finding with the help of a navigation system as its demonstrator. Flexible adaptation of the navigation system to different way-finding goals, situational constraints, and individual differences of the users are described and actually implemented and tested with a prototypical route guidance system.

Error-Induced Learning as a Resource-Adaptive Process in Young and Elderly Individuals, by Ferdinand, Weiten, Mecklinger, and Kray presents neuro-scientific and psychological research in support of Thorndike's hypothesis that human actions followed by positive events are more likely to be repeated than actions followed by negative events (in particular errors), providing recent neuro-scientific evidence for this assumption. The primary aim of their contribution is more specifically to examine under this hypothesis the ability of older adults to flexibly and better adapt to environmental changes, thus compensating for age-related deficiencies. Neural mechanisms underlying error monitoring are introduced and a study examines these processes. Closely related is the work of Kray and Eppinger, which reports about an *ERP-Approach to Study Age Differences in Cognitive Control Processes*. It summarizes the main results of two experiments that aim at the investigation of interactions between two types of cognitive control processes – the ability to switch between task sets and to inhibit well-learned action tendencies by identifying electrophysiological correlates of control processes during task preparation and task execution. The contribution by von Oertzen, Ghisletta, and Lindenberger is similarly concerned with age studies, offering a more mathematical approach to the *Simulation of Statistical Power in Latent Growth Curve Modelling (LGCM): A Strategy for Evaluating Age-Based Changes in Cognitive Resources.* It presents an engine to assist in the gathering of knowledge about statistical properties of LGCM and related methods through systematic and unconstrained exploration of simulated data. The illustration used cognitive aging studies of 20 epochs (weeks, months, or even years) to show that the power to detect variances of changes in an LGCM is dependent on the within-variable level-slope covariance, while the power to detect across-variable covariance of changes in an LGCM is not. The authors conclude that although much cognitive aging literature focuses on change parameters, especially variance and covariance, the field as a whole still does not really know and appreciate the limits of LGCMs.

The final contributions in this section report experimental and computational psycholinguistic research from situated human sentence processing and comprehension. In their contribution, Weber, Crocker, and Knoeferle present a range of experimental findings concerning the processing of *Conflicting Constraints in Resource-Adaptive Language Comprehension*. This work addresses the problem of how the wealth of informational resources which are potentially relevant to situated language comprehension are exploited and prioritized during situated human language processing. Their findings paint a picture in which purely linguistic constraints – long thought to identify the core of sentence comprehension mechanisms – can in fact be overridden by highly contextual aspects of the situation. Mayberry and Crocker outline *The Evolution of a Connectionist Model of Situated Human Language Understanding* which was developed to model the incremental use of visual information

during comprehension. In their architecture, the utterance mediates attention to relevant objects and events in a scene – modeling key eye-tracking findings from Weber et al. – which in turn rapidly influences comprehension. By exploiting a recurrent sigma–pi network that uses an explicit attentional mechanism, their model both demonstrates task-oriented behavior exhibiting and qualitatively models key eye-tracking results observed by Weber et al.

4.2 Part II: Resource-Adaptive Processes in Human–Machine Interaction

The second section of the volume reports work on resource-adaptive processes in human–machine interaction and begins with the *Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech*, by Jameson, Kiefer, Müller, Großmann-Hutter, Wittig, and Rummer. It considers why on-line recognition of resource limitations of the user of a system might be useful in various situations, such as an airline terminal: several experiments with varying noise distraction and time pressure on the user produced typical speech patterns the system had to recognize. The authors trained dynamic Bayesian networks on the resulting data in order to see how well the information in the user's speech could serve as evidence regarding which circumstances the user had been in. The intention is to build more user-adaptive navigation systems and other situated devices based on these findings. The following chapter on the *Shopping Experience of Tomorrow: Human-Centered and Resource-Adaptive* gives an interesting forecast of tomorrow's shopping centers based on RFID-labeled objects and video cameras. This project – widely covered in the media and reminiscent of some scenes in the film "Minority Report" – actually implemented several typical shopping scenarios to research user adaptation in instrumented rooms and user-centered approaches taking the limitations of cognitive and technical resources into account. Wahlster and colleagues demonstrated with several prototypical technical realizations the power of ubiquitous computing and instrumented environments typical for tomorrow's lifestyle. The next contribution "Seamless Resource-Adaptive Navigation" by Schwartz, Stahl, Baus and Wahlster also reports on informed indoor/outdoor navigation. The system's distinguishing features are a ubiquitous, seamless navigation service that is adaptable to the user, the situation the user is in, as well as the technical resources the user has at his or her disposal.

The final contributions to Part II are closely related and lead us into the world of mathematics and proofs in an intelligent learning environment. Wolska, Buckley, Horacek, Kruijff-Korbayová, and Pinkal present a corpus of tutorial dialogs with an ITS to demonstrate *Linguistic Processing in a Mathematics Tutoring System* and specifically examine *Cooperative Input Interpretation and Dialogue Modeling*. The interesting mix of natural language processing interspersed with mathematical formalism is one of the challenges here. Related is Benzmüller, Schiller, and Siekmann's work on *Resource-Bounded Modelling and Analysis of Human-Level*

Interactive Proofs, which takes the results of Wolska et al. as a prerequisite to design and implement a system for intelligent tutoring of mathematical proofs. Correctness, the grain size of the individual proof steps, and automated hint generation are the main challenges in building a tutoring system that can adapt its performance to the individual needs of the student.

4.3 Part III: Resource-Adaptive Rationality in Machines

The final section of this volume deals with resource-adaptive rationality in machines and begins with Wittig's *Comparison of Machine Learning Techniques for Bayesian Networks for User-Adaptive Systems*. Bayesian networks are the central ingredient in several user-adaptive systems presented in this book, and the work reported here provided the networks used by Jameson et al. in particular. In *Scope Underspecification with Tree Descriptions: Theory and Practice*, Koller, Thater, and Pinkal's examine semantic underspecification in natural language processing. They present semantic constructions based on a series of solvers for dominance constraints and dominance graphs, two closely related underspecification formalisms. Debusmann and Kuhlmann continue with natural language processing, presenting their work on grammar formalisms built upon the notion of word-to-word dependencies rather than the more traditional phrase structure grammars. *Dependency Grammar: Classification and Exploration* summarizes the theoretical work of the project on this formalism and shows very competitive results based on its actual implementation. The final chapter in this book is about OMEGA, an automated reasoning system for mathematics. In *Resource-Adaptive Processes in an Automated Reasoning System,* Autexier, Benzmüller, Dietrich, and Siekmann summarize the longstanding research effort to build a mathematical assistance system viewed from the central point of this book: adaptive processes and resource-based computation to cope with the very large search spaces that are typical in this field of research.

Acknowledgments We have freely used material from various sources within the Sonderforschungsbereich, including works from Kuno Lorenz, Werner Tack, and Wolfgang Wahlster. In particular we thank Tony Jameson for permission to adapt his material from the special issue in *Kognitionswissenschaft* vol. 7(3), 1998, Springer Verlag.

References

1. Dean, T.L., Boddy, M. An analysis of time-dependent planning. In: Proceedings of the Seventh National Conference on Artificial Intelligence (pp. 49–54). St. Paul, MN (1988).
2. Gabriel, G. Definitionen und Interessen. Stuttgart- Bad Cannstatt: Frommann-Holzboog (1972).
3. Gigerenzer, G. Adaptive Thinking: Rationality in the Real World. New York: Oxford Union Press (2000).
4. Gigerenzer, G., Selter, R. Bounded Rationality, The Adaptive Toolbox. Cambridge, MA: MIT Press (2001).

5. Hayes-Roth, B. An architecture for adaptive intelligent systems. Journal of Artificial Intelligence, 72(1):329–365 (1995).
6. Jameson, A. Modelling the user's processing resources: Pragmatic simplicity meets psychological complexity. In R. Schäfer, M. Bauer (Hrsg.), Adaptivität und Benutzermodellierung in Interaktiven Softwaresystemen: ABIS'97 (pp. 149–160). Saarbrücken: Universität des Saarlandes (1997).
7. Jameson, A. Ressourcenadaptive kognitive Prozesse. Kognitionswissenschaft, 7(3):95–99 (1998).
8. Kaplan, C.A., Simon, H.A. Foundations of cognitive science. In M.I. Posner (Ed.), Foundations of Cognitive Science (pp. 1–47). Cambridge, MA: The MIT Press (1989).
9. Lenat, D.B. AM: An artificial intelligence approach to discovery in mathematics as heuristic search, Ph.D. Thesis, Stanford University, Stanford, CA (1976).
10. Mittelstraß, J. Die Stunde der Interdisziplinarität? In J. Kocka (Hrsg.), Interdisziplinarität (pp. 152–158). Frankfurt a.M.: Suhrkamp (1987).
11. Norman, D.A., Bobrow, D.G. On data-limited and resource-limited processes. Cognitive Psychology, 7:44–64 (1975).
12. Russell, S.J., Norvig, P. Artificial Intelligence: A Modern Approach. Englewood Cliffs, NJ: Prentice Hall (1995).
13. Russell, S.J., Wefald, E. Do the Right Thing. Studies in Limited Rationality. Cambridge: MIT Press (1991).
14. Schwemmer, O. Handlung und Struktur. Suhrkamp: Frankfurt (1987).
15. Simon, H. Theories of Bounded Rationality. Amsterdam: North Holland (1972).
16. Simon, H. Models of Bounded Rationality. Cambridge, Mass: MIT (1997).
17. Uszkoreit, H. et al. Studien zur performanzorientierten Linguistik. Aspekte der Realtivsatzextraposition im Deutschen. Kognitionswissenschaft, 7(3):129–133 (1998).

Part I
Resource-Bounded Cognitive Processes in Human Information Processing

Visuo-spatial Working Memory as a Limited Resource of Cognitive Processing

Hubert D. Zimmer, Stefan Münzer, and Katja Umla-Runge

1 The Concept of a Resource-Limited Working Memory

Working memory is considered a cognitive component that mainly serves two functions. It *temporarily maintains* information that was either perceived but is no longer present in the environment, or that was internally generated, and it supplies a *work space* for transforming and manipulating elements of perception and thinking. Both functions are relevant for a successful interaction with the environment and it is therefore not surprising that WM is a central topic of research in the field of general psychology. This interest is further increased by the fact that WM is seen as a limited resource that constrains cognitive performances. Understanding working memory capacity (WMC) therefore promises gainful training methods to surmount these capacity limitations. In this chapter, we want to discuss aspects of WM that are relevant when we are talking about WM as a limited resource of cognitive processing. Our focus will be on VSWM although many principles can be applied also to other types of inputs.

Historically, WM research has its origin in research on short-term memory (STM). In 1887 already, Jacobs [47] introduced the so-called digit span to measure the capacity of STM. He presented a random series of digits and participants were required to repeat them in their correct serial order. The longest sequence that was correctly repeated was defined as digit span. It is usually limited to six or seven items and this figure was therefore given as the capacity of STM [76]. Until today, span measures remain the gold standard for estimating WM capacity [18], although nowadays different types of spans are distinguished and a limit of four items is discussed as we will show later. The reason for this differentiation was the observation that STM is not a unitary compartment. Patients with a verbal STM deficit, for example, have a digit span as short as two items, but they have a normal visuo-spatial span [120].

H.D. Zimmer (✉)
Brain and Cognition Unit, Department of Psychology, Saarland University, 66123 Saarbrücken, Germany
e-mail: huzimmer@mx.uni-saarland.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_2,

Those who most clearly made this point and whose model had a strong impact on memory research were Baddeley and Hitch [7]. Because they considered memory as active, they suggested using the term working memory instead of short-term memory. They distinguished three components: a central executive (CE) and two domain-specific independent slave systems – the phonological loop (PL) and the visuo-spatial scratchpad (VSSP) for verbal and visual materials, respectively. According to this model, the CE is associated to attention and it controls the slave systems. The PL stores verbal surface information and it maintains this information by inner speech. In contrast, the VSSP stores visuo-spatial information and a kind of mental inspection was the presumed maintenance mechanism. Additionally, each system has its own limited capacity and therefore selective interference was postulated if two tasks tapping the same system were performed concurrently. If two tasks use the same part system, they compete for resources and dual-task performance is impaired, for example, verbal short-term memory and verbal articulation. If the two tasks are processed in different part systems, performances are similar to a single-task condition, for example, verbal short-term memory and movement tracking [65]. A practical consequence hereof is that one should avoid loading the same system to accomplish two tasks at the same time. Due to the assumed three components, this model is called the tripartite model. It is the only one that features an independent visual working memory. Extended by assumptions on the characteristics of the rehearsal processes, it explains temporary memory for verbal items quite well and it also explains domain-specific differences between memories of verbal and pictorial materials (see [100] for a review). However, it is less successful when dealing with differences *within* the visual domain as we will see in the next section. More recently, the episodic buffer was added as a fourth component of WM [5]. The episodic buffer represents integrated multi-modal information from different systems and modalities including semantic information.

Besides the tripartite model, a number of other suggestions were put forward for explaining WM performances – see the contributions in Miyake and Shah [79]. Some researchers consider (long-term) memory as a network of knowledge entries that can be in a passive or active state, and WM is simply the active part of long-term memory [66]. Capacity limits are also assumed in these models but they are attributed to attention.[1] The number of memory entries that can be in the focus of attention is limited to about four items (e.g. Cowan [20]). Oberauer [80] made an even finer distinction within this set of items. He also assumes that among the active nodes a small set of items is in direct access – this was the formerly mentioned set of four items – but only one item of this set is in the focus of attention. Focused is the item that is selected for a cognitive action, and only this item has a clear processing advantage. The smaller number of only four items compared to the higher digit span of about seven items is due to methodological differences. Complex span measures are enhanced by chunking phenomena – elements can be conjoined – so that a more

[1] A further limitation is given by structural interference. For example, if the items in WM are perceptually similar to each other, memory is worse than if they are different [123].

efficient rehearsal is possible. Four items can be very different things. An item can be a "simple" feature like colour but also a complex unit (a chunk) made of many parts [41]. Hence, if items can be recoded into larger units, more information can be held in WM than if this recoding is not possible. Therefore, training the ability to chunk items is one way to enhance the capacity of working memory [122].

According to the unitary models, WM is not an additional part system or store, but a state of mental processing units, and the main function of WM is providing information for action control. Memory is only a by-product of using information in these processes [27]. Similarly, attention mainly serves this function and the one-element focus of attention is the selection of information for an overt or covert (cognitive) action. These assumptions, however, do not yet explain why only four items are in direct access. This number is empirically well substantiated (see [20]), although there is some variability over individuals and types of to-be-remembered material (see below). The reason for this capacity limit may be found at the neural level. It is very likely that information is represented by synchronised oscillations of cell assemblies. Cell assemblies representing features of the same object oscillate in a synchronous manner, and this synchronicity codes that these features belong to the same object. Simulations of these neural networks showed that only about four items can be stored by this mechanism because neural activities representing more items are not sufficiently separated in time [95]. Limitations due to domain-specific processes are not a main topic in unitary models, although domain-specific storage is conceded and recently its contribution to WM was acknowledged [15, 21]. Hence, WM capacity is probably limited by a domain-general storage mechanism – only a limited number of distinguishable cell assemblies can be simultaneously active – and by domain-specific characteristics of the represented content.

Attention plays also a central role in a third family of WM models of which Engle is one of the leading proponents. According to these models, WMC is related to the ability of controlling central attention [33, 49] and especially to the ability of inhibiting irrelevant information. In support of this assumption, mainly two results are cited. One is the observation that WM capacity correlates with general intelligence and with performances in tasks that have high demands on attentional control (see [32] for a review). The other one is that participants with a low memory span are less efficient in filtering irrelevant background information than those with a high span [17, 31]. In these experiments, span is often defined as "operation span" [116]. In a typical task, participants are sequentially presented with a series of arithmetic equations (e.g. $7 - 3 = 5$) each followed by a word. Participants are required to evaluate the correctness of each equation and to remember the words. After the list, participants have to report the verbal items in their correct order. Hence, a serial verbal memory task is intermixed with arithmetic calculations. This task has high demands on storage and control. It therefore does not surprise that controlled attention is the critical variable that causes the correlation with general intelligence as structural equation models have revealed [34]. More specifically it is executive control, i.e. the ability to allocate attention to the critical task and to resolve task conflicts which distinguish between high-span and low-span participants [99]. In contrast to the control component, the storage component is rather domain-specific

with separate contributions of visual and verbal abilities [50]. A similar picture emerges when visual and verbal mental tests are systematically compared with each other in tasks that have different demands on storage, supervision, and coordination [81, 82].

The many tasks used to investigate WM obviously measure different aspects of working memory and their respective demands determine which component limits memory performances. Therefore, it depends on the used memory paradigm what is considered as WMC. Three independent types of limitations have been identified: (1) memory overload caused by additional perceptual input that enters the same part system and competes for representation (interference), (2) the maximal size of the set of items that can be in direct access, and (3) the efficiency of controlled attention (inhibition of irrelevant items and conflict resolution). Controlled attention seems to be a domain-general ability, whereas as soon as a storage component is involved domain-specific capacities come into play.

2 Components and Capacities of Visual Working Memory

As we have explicated, a domain-specific component contributes to WM and its characteristics depend on the stored content. Two questions therefore arise. What types of information are stored in VSWM and what are the operating characteristics of VSWM? In the model of Baddeley and Hitch, it was assumed that the VSSP represents *spatial* information. This component was not specifically bound to the visual modality because a visuo-spatial main task was impaired by an auditory-spatial secondary task [8]. This view changed some years later when it was observed that presenting irrelevant pictures during maintenance impaired a WM task that makes use of a visual imagery mnemonic [63]. Based on this so-called irrelevant picture effect, Logie concluded that perceived visual information enters the VSSP and interferes with the stored content. More recently, Quinn and colleagues demonstrated that even dynamic visual noise (DVN) – a fast-changing checkerboard randomly filled with black and white dots – interferes with visual imagery [72, 93]. This DVN effect demonstrates that the interference is really a visual effect and not a semantic one. Hence, VSWM should store both types of information and additional spatial as well as visual inputs should impair visuo-spatial memory tasks.

However, the relationships are more complicated. First, the type of the assigned spatial memory task has to be considered. Baddeley and others often used the *Brooks matrix task*. In this task, a spatial sequence of locations within a matrix has to be memorised in correct order (an imaginary path through an empty matrix). The *Corsi task*, which is also frequently used, is structurally very similar. In this task, a sequence of temporarily marked items that are only distinguishable by their spatial location has to be remembered. Active spatial movements impaired the Corsi task [94], as did auditory spatial information [110], and even decision making [51]. It is therefore unlikely that an overload of spatial information in WM by additionally processed spatial information caused the interference. It is more likely that a

disruption of (spatial) attention is critical. In order to maintain a temporal sequence within a set of homogenous elements, spatial marking is necessary, and for that purpose attention is serially directed to locations during the retention interval (spatial rehearsal). Therefore, any process that prevents the allocation of spatial attention during rehearsal should impair spatial memory. We could show that additional spatial processing during maintenance did not interfere with a spatial task if only spatial and not temporal information was relevant [133]. We required our participants to remember the spatial layout of objects. Neither additional visual material nor a spatial suppressor task (spatial tapping) impaired object relocation which of course needs spatial memory. In contrast, performances in a spatio-temporal main task (Corsi) were reduced by spatial but not visual interference. We assumed that directing spatial attention to target locations constitutes spatial rehearsal within VSWM if a spatial sequence in a homogenous field of objects has to be remembered – as in a Corsi task. This rehearsal process is impaired by spatial distraction tasks. In contrast, it seems to be possible to maintain the spatial location of objects by other mechanisms that do not rely on spatial rehearsal as we will discuss later.

In other experiments, the Loci method or the Peg word techniques were used as main tasks. These tasks are imagery mnemonics and they require the generation of a visual image, e.g. imagining a named object at a specific location with the Loci method. Both tasks are impaired by additional visual input [63, 93] and one therefore may assume that imagery mnemonics are generated within VSWM. Considering these differences, Logie [62] suggested to distinguish two components within WM: a visual cache and an inner scribe. The inner scribe operates on the visual cache, it stores dynamic information (processing trajectories of movements and motor actions), and it serves spatial rehearsal. The Corsi task measures the capacity of the inner scribe. The visual cache provides visual information, e.g. shape and colour. Its capacity is measured by the *visual pattern span*. In that task, participants have to remember a pattern of black cells randomly distributed in a matrix for a short time and the number of cells increases from trial to trial [64]. Many results suggest the distinction of different types of visual information (see [62]). However, it remains unclear what type of information is represented in the inner scribe and in the visual cache. For example, spatial information is sometimes investigated as dynamic information and sometimes as configuration of objects (see [133] for a discussion).

We could show that the dynamic characteristic of a visual stimulus is not the critical component for processing of information in the true spatial component (the inner scribe) as it was originally suggested. We investigated WM for biological information (point-light walkers) – a dynamic stimulus – in an S1–S2 task with visual and spatial interference [131]. In the visual interference condition, we presented colour patches during maintenance that should load the visual cache. In the spatial interference condition, spatial tapping was performed during maintenance. Even though in the main task dynamic stimuli were used, we observed visual interference, and this was a function of the similarity between the stimuli of the main and the interference tasks. Irrelevant point-light walkers impaired memory more than irrelevant colours. Spatial tapping also caused interference but this effect was

not due to spatial information because also non-spatial tapping interfered. We therefore assume that the distinction between visual information (appearance) and spatial information – where objects are located – is a critical dimension in VSWM and less the distinction between static and dynamic information. Recently this position was also supported by other studies [25].

This separation between object and location information in WM follows the separation of these two features in perception (e.g. [119]). It has been demonstrated both in behavioural and neurocognitive studies which we will discuss later. With this definition, however, spatial information is no longer confined to visual input because also other objects, e.g. sound objects, are perceived at specific locations. We tested the idea of common spatial coding in WM in a series of experiments on auditory spatial and visuo-spatial location memory. Our data clearly speak in favour of a common coding in spatial WM (SWM) independent of modality (Fig. 1).

We presented lists composed of only visual or of visual and auditory material in a spatial working memory task. If modality-specific SWMs exist each should have its own capacity. Because the capacity of each system limits the amount of remembered items, memory performance should be higher if the available capacity is higher. Hence, performance should be higher if memory load can be distributed over two memories (mixed-modality list) than if the complete list has to be hold within one component (uni-modality list). In contrast, if all spatial information is stored within one system, performances should be a function of list length independent of the modality of the items. We therefore presented uni- and mixed-modality

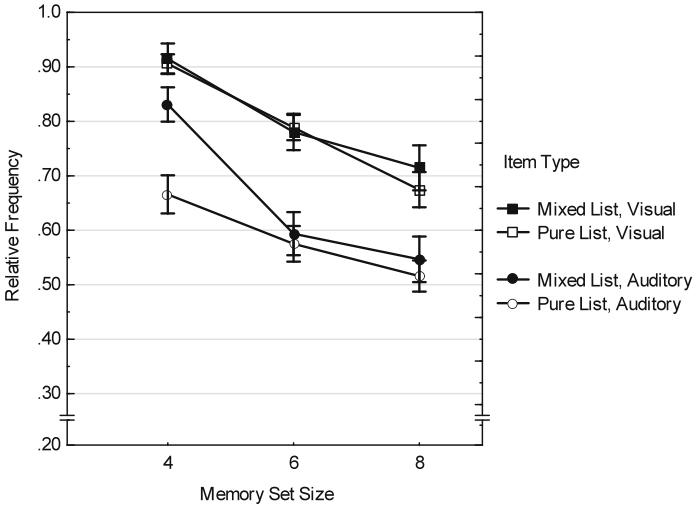


Fig. 1 Memory performances in a visuo-spatial working memory task as a function of modality (visual, auditory) and list type (pure, mixed), data from [60]. Note that the advantage for auditory material in a four-item mixed list is not a memory advantage. It is due to an advantage in auditory spatial perception caused by reduced spatial uncertainty (see the additional data presented in the original work)

lists of visual and auditory objects within a S1–S2 matching task framework [60]. Memory load was a function of the number of items and it was not dependent on input modality. Auditory-spatial and visuo-spatial tasks seem to be processed within the same working memory system.

We ran comparable experiments as electrophysiological studies and these data supported the same conclusion. Event-related potentials during encoding were influenced by modality of the stimuli, but slow potentials during maintenance were a function of load but not of modality [61]. We therefore concluded that spatial information of visual and auditory input is processed and maintained by the same SWM. The spatial system probably represents object locations in egocentric coordinates relative to the coordinates of the observer's head.[2] In contrast, object-specific information is represented in separate domain-specific components according to its content, e.g. visual or auditory representations within a visual WM (VWM) and an auditory WM (AWM), respectively [3, 96].

These domain-specific components have their own constraints. The VWM can represent only a limited number of objects simultaneously. Luck and Vogel [67, 123] estimated the capacity of VSWM as about three to four items – they made no distinction between visual and spatial information. They presented their participants objects (e.g. colour patches or conjunctions of colours and shapes) and after a short delay a test picture was presented. The number of objects was varied. In all conditions, memory was nearly perfect up to four objects (features or conjunctions of features) and then it declined. The authors concluded that VWM can hold four objects and that each object can represent an arbitrary number of features without any additional costs. Hence, VWM is limited in the number of objects not in the number of features. The capacity limit of VWM was therefore set to four objects. Meanwhile, however, we know that feature conjunctions cause additional processing effort [114] and that the capacity also depends on physical characteristics of the "objects". Figures have to be spatially coherent to define an object [126], and the more visually complex items are (e.g. Chinese letters vs. colour squares) the less items that can be remembered [1]. Visual working memory is therefore not only limited by the number of items of the same type, but also by the structural qualities of these items.

However, even the assumption that VSWM consists of two independent part systems – representing visual appearance and spatial information – has to be further differentiated. A frequently used WM task is the S1–S2 matching task. A to-be-remembered stimulus is presented and shortly afterwards either the same or a changed stimulus is presented for comparison. In a dual-task condition, additional material has to be processed in as secondary task. If the second task loads the same WM component as the main task, performances are reduced compared to the single-task condition. It is assumed that both tasks together exceed the capacity of WM

[2] We will not go into details of spatial representations. A closer look would reveal different types of spatial relations, for example, ego-centric and allocentric spatial representations with different reference systems [126].

and therefore performances collapse. However, when the additional information is exclusively presented during maintenance, often no interference was observed [2, 4, 91, 132]. Even demanding secondary tasks during maintenance did not influence visuo-spatial memory in a S1–S2 working memory task [130]. This should not happen if the additionally perceived material automatically enters WM and competes with the material of the main task. Logie [62] explained the absence of interference by the assumption that visual input does not automatically and directly enter WM but it does so only after some kind of higher cognitive processing that filters visual input. Consequently, only relevant information would enter WM, whereas irrelevant information would be inhibited although it is perceived. However, this explanation does not work either because other working memory conditions that are impaired by additional (irrelevant) input during maintenance do exist. For example, dynamic visual noise during maintenance impaired a visual, just noticeable difference size task [71] – in this task the size of a circle has to be compared with another one presented a few seconds before. Similarly, we observed that temporary visual memory for movements (point-light walkers) was impaired by additional visual input (colour patches or other walkers to be ignored) [131]. Quinn and colleagues therefore argued to distinguish between passively hold visual material and material that is held in an active visual buffer [90, 91]. According to their view, visual input has direct access to a visual store [70, 92], but this visual store is not the passive visual cache but an active visual buffer [90, 91]. This buffer holds conscious visual representations of perceived stimuli as well as of mentally generated ones (visual images).

Within the visual working memory framework, a separate buffer was already postulated by Pearson [84]. This buffer should hold a conscious visual image, and it should therefore be closely tied to central executive and conscious awareness. The idea of a *visual buffer* that is used for image generation was originally put forward by Kosslyn and a lot of research in the context of the so-called imagery debate was devoted to this structure and its characteristics [55–57]. The controversy was about the quasi-pictorial quality of representations within this buffer. Kosslyn assumed that the buffer represents visual information in a depictional manner, preserving distance between represented objects, and all processes on this buffer are analogous to processes in physical space. Mental rotation for example is a stepwise process passing intermediate positions between the start and the target orientation [107], and mental scanning follows a trajectory in a two-dimensional space with Euclidean characteristics [58]. However, until today it is controversial whether these characteristics are caused by constraints of the mental (or neural) representation or they are only simulations of the real world [89]. Nevertheless, many results have shown that imagery processes, for whatever reason, behave as if they were performed in physical reality. For example, visual images “have” a specific resolution, size, include angles and distances between objects, etc. and these features follow the same processing characteristics as their physical counterparts [55]. One can therefore consider these features as quasi-physical constraints of visual imaginal processing. They come into effect when information is processed within the visual buffer.

In summary, the conception of VSWM is more differentiated than in its beginning. WM for spatial information is distinguished from WM for objects. SWM is closely related to spatial attention and rehearsal is provided by shifts of spatial attention. Objects are represented in domain-specific WM in the case of visually perceived objects, very likely as distributed representations of visual features (see below). The maximal number of objects is limited to about four items. Additionally, however, the complexity of stimulus material restricts the number of successfully remembered items. It may be that both limitations are caused by different mechanisms. Furthermore, we have to distinguish passive and active temporary memories of visual information. The former is not actively maintained, whereas the latter one is held active in a visual buffer and is closely related to attention.

3 Working Memory and Higher Cognitive Performances

In the classical, structural view on WM, domain-specific storage components and executive functioning – which can be seen as higher cognitive processes – are viewed as separate mechanisms, or resources, and consequently, they are measured separately [6]. For instance, the storage capacity of verbal WM was measured with simple span tasks (e.g. with classical digit span in the verbal domain) and the capacity of VSWM with the Corsi blocks task [77]. Another example is the arrow span task that requires remembering a series of directions successively indicated by an arrow. Examples for measurements of the central executive function are random number generation (asking a participant to generate a sequence of numbers that have a random order) and the Tower of Hanoi problem (minimising the number of moves, participants have to rebuild a tower of discs considering several constraints).

Other scientists focused on simultaneous storage-and-processing tasks to measure the capacity of working memory (e.g. [23]). According to this approach, WM performance is measured as the number of elements that can be remembered in the face of ongoing processing. A typical task requires a participant to process some information (e.g. to read a sentence aloud or to evaluate whether a simple mathematical equation is correct) and simultaneously store some information besides processing (e.g. remember the last words of the preceding sentences or remember the results of the preceding equations). The number of to-be-remembered items that can be recalled after a list of such sentences or equations has been processed is the individual working memory span. Individuals differ reliably in such storage-and-processing tasks (see [18] for a review). Classical measures are reading span [23], operation span [116], and counting span [14]. In the visuo-spatial domain, similar processing-and-storage measures have been constructed. In the rotation letter span [106], a picture of a rotated letter is shown. This rotation is produced either with the original letter or with a mirrored version of the letter. Participants are asked to judge whether the letter is mirrored or not (processing component) while remembering the directions of the rotations of preceding letters, similar to the arrow span task (storage component).

These storage-and-processing parameters have become particularly important when one is interested in higher cognitive performances because they have a diagnostic value for higher cognitive tasks and intelligence. Studies investigating the relation of WM to cognitive performances in different domains and application fields (e.g. language comprehension, spatial ability, and environmental learning) typically utilise an individual differences approach. They relate the individual WM capacity measured by a specific span to performances in higher cognitive tasks. For example, Daneman and Merikle conducted a meta-analysis of 77 studies in which the association between WM capacity and language comprehension ability was investigated [24]. WMC as measured with storage-and-processing tasks was a good predictor of language comprehension. The predictive power of such tasks was higher than the predictive power of STM tasks that measure storage alone. Moreover, not only reading span (working memory and language processing) but also operation span (working memory and math processing) predicted language comprehension.

In the context of storage-and-processing tasks, the issue of whether there are domain-specific WMCs is controversial. Shah and Miyake [106] propose the idea that there are different domain-specific resource pools that fuel two domains of higher-level cognition: spatial thinking and language comprehension. In their study, they related visuo-spatial (letter rotation span and arrow span) and verbal WMC (reading span) to psychometric tests of spatial ability and to scores of language ability. Three spatial visualisation tests (Paper Form Board Test, Space Relations Test, and Clocks Test) and one perceptual speed test (Identical Pictures Test) were used. The Paper Form Board Test, for example, consists of several drawings of two-dimensional pieces that can be put together. A clear pattern was obtained. The visuo-spatial WMC measure (letter rotation span) was strongly related to a composite measure of the spatial visualisation tests, whereas it was not related to verbal ability. Vice versa, verbal WMC (reading span) was considerably related to verbal ability but not to spatial ability [106]. In another study, it was observed that performances in the Tower of Hanoi task were related to spatial span but not to verbal span measures, whereas conditional reasoning was related to the reading span task but not to the spatial span task [42]. Thus, not only storage functions as investigated in S1–S2 tasks, but also higher-level functions and problem-solving tasks involving the central executive appear separable with respect to domain-specificity and further to visual and verbal WM capacities.

Recently, attempts have been made to structure WM functions more theoretically, in order to clarify their role in cognitive performance and intelligence. As in models of intelligence, Oberauer and colleagues [81] have differentiated the working memory construct along the content dimension (verbal, spatial-figural, and numerical tasks) and the function dimension (storage-and-processing, supervision, and coordination). In order to test this, their participants performed a series of WM tasks with different cognitive demands according to these two dimensions. The studies however yielded mixed outcomes for the separation of different WM components. The results spoke predominantly in favour of the assumption that spatial components can be separated from verbal ones, but there was a less clear separation of processes.

Finally, visuo-spatial WMC was related to *visuo-spatial abilities* as measured in intelligence tests. Spatial abilities are traditionally measured by paper-and-pencil tests, and they represent the spatial dimension of intelligence. For example, these tests require to mentally rotate a three-dimensional object, to find an embedded figure, or to mentally fold and unfold a piece of paper. In their investigation of the relation between VSWM and spatial abilities, Miyake and colleagues, e.g. [78], have included storage-and-processing tasks (e.g. the letter rotation task), short-term storage tasks (e.g. the Corsi blocks task), and central executive tasks (e.g. the Tower of Hanoi). It has been found that different spatial ability tasks put different demands on WM. There are tasks of spatial visualisation which "reflect processes of apprehending, encoding, and mentally manipulating spatial forms" ([13], p. 309). These tests require to perform a series of transformations on mental representations of objects (such as mentally folding a piece of paper), and they appear closely related to central executive functioning. However, tests that require rather low mental manipulation (such as identifying a picture in a row of similar pictures, classified as perceptual speed) are more directly related to VSWM. Moreover, in the visuo-spatial domain, the storage-and-processing tasks and the storage tasks appear to measure the same construct [78]. This observation is to be expected when we consider the fact that the visual buffer as a work space for active maintenance and imaginal processing is closely tied to conscious awareness and the central executive.

4 Neural Structures Underlying Working Memory

The different WM functions are provided by a distributed network of active brain structures. Neurocognitive studies have shown that anterior and posterior cortical structures contribute to WM. These are mainly the dorsolateral and ventrolateral prefrontal cortex (referred to as DLPFC and VLPFC in the following) on the one hand and distributed regions in occipital, parietal, and temporal cortex on the other. Functionally, posterior structures have been ascribed the role of passive temporary buffers specialised for different representation formats. In contrast, anterior structures are assumed to be more active modules providing rehearsal and manipulation mechanisms for the contents of working memory as well as monitoring the posterior subsystems in a central executive manner.

The results of the first neuroimaging studies on WM indicated a hemispheric specialisation. Verbal WM seemed to rely primarily on a left-hemispheric network whereas visuo-spatial information in WM required mainly cortical structures in the right hemisphere [109]. While for verbal WM the specialisation of the left hemisphere has been repeatedly demonstrated and is therefore widely accepted, the right hemisphere's dominance in short-term retention of visuo-spatial material has been questioned [124].

In contrast to the less stable right lateralisation of VWM, the specialisation of dorsal and ventral areas for processing of spatial and visual information respectively is empirically well supported. A major result from a range of neurocognitive studies

is the dissociation of posterior brain structures contributing to visual object (e.g. shape, colour, and texture) and spatial processing, sometimes also called the "what" and "where" processing streams. Occipito-temporal regions showed more activation when object features of visual stimuli were to be retained whereas spatial features in WM led to stronger activations in occipito-parietal structures [9, 19, 73, 101, 121]. By transcranial magnetic stimulation (TMS) it was further demonstrated that these areas are in fact functional for WM tasks and their activity is not an epiphenomenon. In this paradigm, repetitive pulses through a coil (shaped as an eight) that is placed over the brain area of interest temporarily disturb processing in the respective structures. Ranganath and colleagues [98] identified regions in the fusiform face area and in parahippocampal place area that show category-specific activity. When these neural structures were temporarily interrupted by TMS, a stimulation of the temporal cortex slowed responses in object tasks, whereas stimulating the parietal cortex caused slower responses in spatial tasks [54].

If we look at posterior structures involved in WM retention at feature level, we can state that these processing structures can be further subdivided. In functional magnetic resonance imaging (fMRI) studies, feature-specific brain structures were found active dependent on the type of information that was maintained. During short-term retention of object categories, specific brain areas were preferentially active: fusiform, lingual, and inferior temporal cortex for shape information [88, 121], posterior parietal cortex including intraparietal sulcus for positions [22], fusiform face area for facial identity [29, 87], and parahippocampal place area for places and scenes [97, 98].

In a recent fMRI study from our lab, we directly compared movement, position, and colour information during retention in working memory [117]. Participants' working memory for selective features of dynamic and static stimuli was tested within an S1-cue-S2 paradigm consisting of two coloured dots (S1 and S2). The cue indicated the feature that had to be compared and it was presented after S1 offset to focus on selective rehearsal and to circumvent effects of selective encoding. A retention interval of 6,000 ms followed, in that the respective information should be rehearsed. We contrasted brain activity during this maintenance interval as a function of the cued feature. For dynamic stimuli, either movement or end-position information and for static stimuli, either position or colour information had to be retained in working memory. Results indicate that regions that are also involved in movement perception (MT/V5, superior temporal sulcus, and premotor cortex) were differentially activated during short-term retention of movement information. Position WM (with static or dynamic stimuli) especially recruited parahippocampal regions and the lateral occipital complex (LOC), structures known to be involved with spatial representations of objects. Furthermore, left fusiform cortex (a structure belonging to the ventral processing stream) was significantly more activated when participants retained the coloured dot's end position in working memory as compared to its movement. This result (together with parahippocampal and LOC activations) strengthens the perspective that a coloured dot's position is remembered as an object at a specific location having specific features (e.g. colour). In contrast, movement information can be rehearsed in a more abstract way without

the necessity to imagine the moving object with its features. Selective retention of colour information in WM yielded activations in the anterior portion of the right superior temporal gyrus and in early visual processing regions. This suggests that for short-term retention feature-selective posterior regions can be defined which in part comprise the same structures that are active during domain-specific perception (Fig. 2).

While posterior structures show a domain-specific organisation, the functional role and specificity of regions in the prefrontal cortex is controversial. Two main perspectives can be identified: the *domain-specific hypothesis* and the *process-specific hypothesis*. The first perspective assumes that the "what" and "where" dissociation equally holds for prefrontal cortex [118]. Short-term retention of spatial information involves DLPFC whereas VLPFC is mainly concerned with the retention of visual object information. An alternative view is the dissociation of prefrontal cortex as to the processes that are applied to information in WM [83, 85, 86]. According to this perspective, VLPFC is involved in active maintenance of to-be-remembered stimuli whereas DLPFC deals with the manipulation of working memory content. Although this controversy is not resolved, it is likely that the PFC is not the storage site of sensory information. These regions may function as pointers to posterior feature-specific areas and therefore they appear as if they represent domain-specific information. We assume that anterior and posterior areas are part of the network that provides VSWM [36, 103]. The anterior structures refer to posterior domain-specific representations and keep them active during maintenance. The domain-specific brain areas are widely the same that represent the visual features in perception and these structures also function as storage sites for both working and long-term memory processes [127].

Several of these brain areas found active during WM maintenance are also involved in visuo-spatial reasoning and in visual imagery tasks. In an fMRI study,

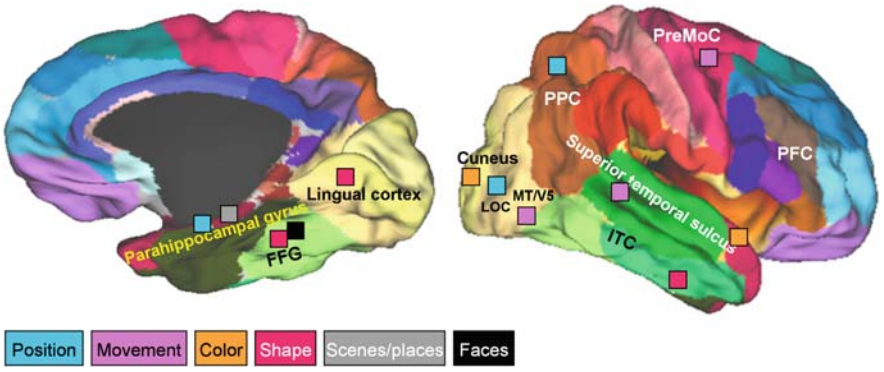


Fig. 2 An illustration of brain structures that were found active in different working memory tasks – *left*: medial view, *right*: lateral view on right hemisphere. *Coloured squares* indicate the type of feature that was relevant in the task. PFC, prefrontal cortex; PPC, posterior parietal cortex; PreMoC, premotor cortex; ITC, inferior temporal cortex; LOC, lateral occipital cortex; FFG fusiform gyrus

Todd and Marois [113] observed that activity in the intraparietal sulcus (IPS) and intraoccipital sulcus increased with the number of items in WM and the signal change correlated with the individual WM capacity [112]. Activity in the IPS was correlated to memory load and task difficulty [111]. Furthermore, the IPS was active in a mental scanning task [74] and in different types of spatial imagery tasks [69]. Also a spatial reasoning task showed enhanced activity in the IPS [115] and TMS of this structure impaired performances [105]. During mental rotation, negative slow potentials were observed at parietal electrodes [46] that increased with angular disparity [102]. Activity of the parietal cortex was also found in an fMRI study when figures, letters, or abstract shapes were rotated [48]. In contrast, object imagery tasks caused enhanced activity in the occipito-temporal and inferior-temporal cortex [28, 40, 75]. Kosslyn argued that even areas in the occipital cortex that are involved in early processing of visual input are activated if detailed visual images were generated [59]. Taken these and several other studies, it is likely that at the neural level the separation of visual and spatial processing also exists in imagery tasks and that occipito-temporal and parietal structures process these types of information, respectively. However, we only begin to understand the neural network that provides these imagery processes.

5 Visual Working Memory in an Applied Context

We have already presented a number of cognitive tasks that were influenced by VSWM efficiency but most of these tasks were artificial. They were designed to provide a relatively pure measure of VSWM in the laboratory. In this chapter we will have a look at visuo-spatial tasks that are relevant in applied settings.

Environmental learning. It appears that in the visuo-spatial domain specific task requirements constrain what resources are employed for cognitive performance. For instance, spatial abilities were related to spatial layout learning only if the learning experience was solely visual [43, 125]. In contrast, if the environment was studied by direct experience, then spatial abilities played a minor role. Hegarty and Waller [45] have therefore suggested that spatial abilities as measured in the laboratory – usually transformations of objects – should be distinguished from environmental spatial abilities. The latter, but not the former, require integrating different views of the environment over time to form a coherent mental representation. Moreover, the spatial reference frames differ. Navigational experience corresponds to the egocentric reference frame, which is view-based, depending on the current position and orientation in the environment. The orientation changes with one's own movements and spatial configurations are coded in relation to the body axes. Spatial ability tests, in contrast, require the comparison and/or mental manipulation of objects that can be apprehended in a single view. The reference frame here is allocentric, i.e. the spatial properties of the object are related to a fixed external coordinate system. Similarly, environmental survey knowledge (like on a map) corresponds to the allocentric

reference frame, which is independent of the individual's position in the environment. However, a map is so complex that learning is based on fragments or regions that are put together mentally [128]. In support of this distinction, the relationship between spatial ability tasks and spatial performance tests in the environment was rather weak (see [45]). Bosco and colleagues related four different VSWM tasks to spatial orientation tasks measuring landmark knowledge, survey knowledge, and route knowledge [10]. VSWM spans explained only a limited percentage of the variance in the orientation tasks. However, all tasks were administered in the laboratory, i.e. they were solely visual.

When learning about the spatial configuration of an environment is based on direct experience, memory representations of different parts of the environment have to be maintained and integrated. Thus, WM may play an important role in spatial *environmental* learning [43]. Garden et al. [37] demonstrated that secondary tasks had detrimental effects on route learning in a real environment. Route learning was realised by following the experimenter through the centre of Padua. High-spatial ability participants appeared more affected by a spatial interference task, whereas low-spatial ability participants were more affected by a verbal interference task. Hegarty and colleagues have found considerable correlations between a VSWM measure (the arrow span) and learning from direct experience in the environment [43]. Similarly, in a series of experiments involving learning during navigation through a novel, real-world environment, we have found a substantial and robust relation between VSWM capacity and the acquisition of orientation knowledge in a real environment (see Zimmer et al., Visuo-Spatial Working Memory as a Limited Resource of Cognitive Processing of this volume).

In summary, the capacity of VSWM appears as a critical resource in real-world spatial orientation tasks. However, the functional role of VSWM in environmental learning has not been studied systematically yet.

Learning from diagrams and animations. When a static diagram of a mechanical system is studied, the movement of the system components has to be inferred by some mental processing. This mental processing has been characterised as involving "envisioning", "running a mental model", or "simulating the behaviour of a system in the mind's eye" (e.g. [38]). The term "mental model" refers to a mental representation that is dynamic and spatial. Such a representation allows to mentally simulate the system's operation. Dual-task studies with either verbal or visuo-spatial load while trying to understand the movement of the components of a simple mechanical system have shown that VSWM is involved [108]. Utilising an individual differences approach it was demonstrated that participants with low spatial visualisation ability (as measured with the Paper Folding Test, the Vandenberg Mental Rotation Test, the Guilford–Zimmerman Spatial Orientation Test, and the Bennett Mechanical Comprehension Test) performed worse on diagram comprehension and mental animation [44]. Thus the rather vague idea of "running a mental model" appears closely related to spatial visualisation and VSWM memory resources. It is likely that the efficiency of visual imagery causes this dependence.

These resources might also be critical if dynamic visual animations are presented in some multimedia learning material, that is, if mental animation is unnecessary because the movement of the system components is shown explicitly. Mayer and Sims [68] found that high-spatial ability subjects benefited from the concurrent presentation of an animation of a mechanical system with a narration as compared to a successive presentation (contiguity effect), while low-spatial ability subjects did not benefit from the concurrent presentation [68]. This result was explained with the ease with which high-spatial ability subjects could understand the explicit animation. Because of this superiority, more central WM resources could be devoted to building relations between the verbal narration and the visual animation. Low-spatial ability subjects, in contrast, had to devote central working memory resources to the understanding of the animation itself. In a task involving understanding of the spatial structure of a complex 3D-object, Cohen and Hegarty [16] provided subjects with interactive control over explicit animations. Large individual differences were found with respect to effectiveness of use of these interactive animations. High-spatial ability subjects were more likely to use the external animations. Thus, there was no evidence that low-spatial ability subjects used the external animations to compensate for poor mental animation. This result suggests that external multimodal support of learning needs additional assistance to enhance performances of users with low visuo-spatial abilities.

Spatial reasoning. Many of the WM tasks that we presented made spatial inferences necessary. Spatial reasoning can therefore be considered a prototypical task of SWM. However, many additional experiments exist that were specifically designed to investigate more complex spatial reasoning tasks. Already Brooks [11, 12] demonstrated that simultaneous processing of spatial relations and visually presented sentences interfered with each other and caused lower performances than auditory input. Similarly, Glass and Eddy reported better performances with auditory than with visual presentation of sentences if visual features were verified [30, 39]. They required their participants to ask questions on spatial relations between features of objects. These results can be explained by assuming that visually presented sentences and processing of visual features cause work load within the same WM component whereas auditory-verbal processing does not. Compatible with such an assumption we found evidence for visual processing if "inspection" of a visual image was needed in order to verify relations between stimuli but not if the answer could be retrieved from abstract propositional knowledge [129]. The fMRI studies reported above suggest that these tasks are processed in neural structures that are dedicated to visuo-spatial processing. A direct test of the involvement of the WM structures in reasoning was presented by Knauff and colleagues. While participants solved reasoning tasks with spatial relations, activity in the occipital parietal cortex was observed [52, 53, 104]. Interestingly, if visual imagery was used to solve these tasks the quasi-analogous features of visual images were also effective. Decision times in mental scanning were a function of Euclidean distance even if the mental map was constructed from texts [26]. Decision times were a function of the degree of mental rotation and they correlated with the hemodynamic response [35].

Obviously, solving visuo-spatial problems induces visual-imaginal processing. This type of processing is accompanied on the one hand with activations in spe-

cific neural structures and on the other hand with specific processing characteristics analogous to physical processes. Future research has to show whether the physical analogue characteristics are constraints of the neural structures or of the participants' experience with the physical world. However, independent of the answer to this question, the reported results at the behavioural and neural level already provide evidence for a separate VSWM processing resource. This part system processes a specific type of information, it has a specific capacity, it has specific processing characteristics, and it is provided by specific neural structures.

References

1. Alvarez, G.A., Cavanagh, P. The capacity of visual short term memory is set both by visual information load and by number of objects. Psychological Science, 15:106–111 (2004).
2. Andrade, J., Kemps, E., Werniers, Y. Insensitivity of visual short-term memory to irrelevant visual information. Quarterly Journal of Experimental Psychology, 55A: 753–774 (2002).
3. Arnott, S.R., Grady, C.L., Hevenor, S.J., Graham, S., Alain, C. The functional organization of auditory working memory as revealed by fMRI. Journal of Cognitive Neuroscience, 17: 819–831 (2005).
4. Avons, S.E., Sestieri, C. Dynamic visual noise: No interference with visual short-term memory or the construction of visual images. The European Journal of Cognitive Psychology, 17:405–424 (2005).
5. Baddeley, A.D. The episodic buffer: A new component of working memory? Trends in Cognitive Sciences 4:417–423 (2000).
6. Baddeley, A.D. Working Memory. Oxford: Oxford University (1986).
7. Baddeley, A.D., Hitch, G. Working memory. In G.A. Bower (Ed.), Recent Advances in Learning and Motivation (vol. 8, pp. 47–90). New York: Academic Press (1974).
8. Baddeley, A.D., Lieberman, K. Spatial working memory. In R. Nickerson (Ed.), Attention and Performance VIII (pp. 521–539). Hillsdale: Lawrence Erlbaum (1980).
9. Bosch, V., Mecklinger, A., Friederici, A.D. Slow cortical potentials during retention of object, spatial, and verbal information. Cognitive Brain Research, 10:219–237 (2001).
10. Bosco, A., Longoni, A.M., Vecchi, T. Gender effects in spatial orientation: Cognitive profiles and mental strategies. Applied Cognitive Psychology, 18:519–532 (2004).
11. Brooks, L.R. Spatial and verbal components of the act of recall. Canadian Journal of Psychology, 22:349–368 (1968).
12. Brooks, L.R. The suppression of visualization by reading. Quarterly Journal of Experimental Psychology, 19:289–299 (1967).
13. Carroll, J.B., Frederiksen, N., Mislevy, R.J., Bejar, I.I. Test Theory and the Behavioral Scaling of Test Performance. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. (1993).
14. Case, R., Kurland, D.M., Goldberg, J. Operational efficiency and the growth of short-term memory span. Journal of Experimental Child Psychology, 33:386–404 (1982).
15. Chen, Z., Cowan, N. Chunk limits and length limits in immediate recall: A reconciliation. Journal of Experimental Psychology: Learning, 31:1235–1249 (2005).
16. Cohen, C.A., Hegarty, M. Individual differences in use of external visualisations to perform an internal visualisation task. Applied Cognitive Psychology, 21:701–711 (2007).
17. Conway, A.R.A., Cowan, N., Bunting, M.F. The cocktail party phenomenon revisited: The importance of working memory capacity. Psychonomic Bulletin & Review, 8:331–335 (2001).
18. Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., Engle, R.W. Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin & Review, 12:769–786 (2005).

19. Courtney, S.M., Ungerleider, L.G., Keil, K., Haxby, J.V. Object and spatial visual working memory activate separate neural systems in human cortex. Cerebral Cortex, 6:39–49 (1996).
20. Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24:87–185 (2001).
21. Cowan, N., Morey, C.C. How can dual-task working memory retention limits be investigated? Psychological Science 18:686–688 (2007).
22. Curtis, C.E. Prefrontal and parietal contributions to spatial working memory. Neuroscience 139:173–180 (2006).
23. Daneman, M., Carpenter, P.A. Individual differences in working memory and reading. Journal of Verbal Learning and Verbral Behavior, 19:450–466 (1980).
24. Daneman, M., Merikle, P.M. Working memory and language comprehension: A metaanalysis. Psychonomic Bulletin & Review, 3:422–434 (1996).
25. Darling, S., Della Sala, S., Logie, R.H. Behavioural evidence for separating components within visuo-spatial working memory. Cognitive Processing, 8:175–181 (2007).
26. Denis, M., Zimmer, H.D. Analog properties of cognitive maps constructed from verbal descriptions. Psychological Research, 54:286–298 (1992).
27. D'Esposito, M. From cognitive to neural models of working memory. Philosophical Transactions of the Royal Society B, 362:761–772 (2007).
28. D'Esposito, M., Detre, J.A., Aguirre, G.K., Stallcup, M., Alsop, D.C., Tippet, L.J., Farah, M.J. A functional MRI study of mental image generation. Neuropsychologia 35:725–730 (1997).
29. Druzgal, T.J., D'Esposito, M. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. Journal of Cognitive Neuroscience, 15:771–784 (2003).
30. Eddy, J.K., Glass, A.L. Reading and listening to high and low imagery sentences. Journal of Verbal Learning and Verbral Behavior, 20:333–345 (1981).
31. Elliott, E.M., Barrilleaux, K.M., Cowan, N. Individual differences in the ability to avoid distracting sounds. European Journal of Cognitive Psychology, 18:90–108 (2006).
32. Engle, R.W., Kane, M.J., Ross, B.H. Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B.H. Ross (Ed.), The Psychology of Learning and Motivation (vol. 44, pp. 145–199). New York: Elsevier Science (2004).
33. Engle, R.W., Kane, M.J., Tuholski, S.W. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In P.S. Akira Miyake (Ed.), Models of Working Memory (pp. 102–134). Cambridge: Cambridge University Press (1999).
34. Engle, R.W., Tuholski, S.W., Laughlin, J.E., Conway, A.R.A. Working memory, short-term memory, and general fluid intelligence: A latent variable approach. Journal of Experimental Psychology: General, 128:309–331 (1999).
35. Formisano, E., Linden, D.E., Di Salle, F., Trojano, L., Esposito, F., Sack, A.T., Grossi, D., Zanella, F.E., Goebel, R. Tracking the mind's image in the brain I: Time-resolved fMRI during visuospatial mental imagery. Neuron, 35:185–194 (2002).
36. Fuster, J.M. Network memory. Trends in Neurosciences, 20:451–459 (1997).
37. Garden, S., Cornoldi, C., Logie, R.H. Visuo-spatial working memory in navigation. Applied Cognitive Psychology, 16:35–50 (2002).
38. Gentner, D., Stevens, A. Mental Models. Hillsdale: Lawrence Erlbaum, (1983).
39. Glass, A.L., Eddy, J.K., Schwanenflugel, P.J. The verification of high and low imagery sentences. Journal of Experimental Psychology [Human Learning] 6:692–704 (1980).
40. Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., Singer, W. The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. European Journal of Neuroscience, 10:1563–1573 (1998).
41. Halford, G.S., Cowan, N., Andrews, G. Separating cognitive capacity from knowledge: A new hypothesis. Trends in Cognitive Science, 11:236–242 (2007).
42. Handley, S.J., Capon, A., Copp, C., Harper, C. Conditional reasoning and the Tower of Hanoi: The role of spatial and verbal working memory. British Journal of Psychology, 93:501 (2002).

43. Hegarty, M., Montello, D.R., Richardson, A.E., Ishikawa, T., Lovelace, K. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. Intelligence, 34:151–176 (2006).
44. Hegarty, M., Sims, V.K. Individual differences in mental animation during mechanical reasoning. Memory & Cognition, 22:411–430 (1994).
45. Hegarty, M., Waller, D.A. (Eds.). Individual Differences in Spatial Abilities. New York: Cambridge University Press (2005).
46. Heil, M., Bajric, J., Rösler, F., Hennighausen, E. Event-related potentials during mental rotation: Disentangling the contributions of character classification and image transformation. Journal of Psychophysiology, 10:(1996) 326–335.
47. Jacobs, L.L. Experiments in "prehension". Mind, 12:75–79 (1887).
48. Jordan, K., Heinze, H.J., Lutz, K., Kanowski, M., Jäncke, L. Cortical activations during the mental rotation of different visual objects. Neuroimage, 13:143–152 (2001).
49. Kane, M.J., Bleckley, M.K., Conway, A.R.A., Engle, R.W. A controlled-attention view of working-memory capacity. Journal of Experimental Psychology: General, 130:169–183 (2001).
50. Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., Engle, R.W. The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology: General, 133:189–217 (2004).
51. Klauer, K.C., Stegmaier, R. Interference in immediate spatial memory: Shifts of spatial attention or central-executive involvement? Quarterly Journal of Experimental Psychology A, 50:79–99 (1997).
52. Knauff, M., Fangmeier, T., Ruff, C.C., Johnson-Laird, P.N. Reasoning, models, and images: Behavioral measures and cortical activity. Journal of Cognitive Neuroscience, 15:559–573 (2003).
53. Knauff, M., Mulack, T., Kassubek, J., Salih, H.R., Greenlee, M.W. Spatial imagery in deductive reasoning: A functional MRI study. Brain Research: Cognitive Brain Research, 13: 203–212 (2002).
54. Koch, G., Oliveri, M., Torriero, S., Carlesimo, G.A., Turriziani, P., Caltagirone, C. rTMS evidence of different delay and decision processes in a fronto-parietal neuronal network activated during spatial working memory. Neuroimage, 24:34–39 (2005).
55. Kosslyn, S.M. Mental images and the brain. Cognitive Neuropsychology, 22:333–347 (2005).
56. Kosslyn, S.M. Image and brain. The Resolution of the Imagery Debate. Cambridge: MIT Press (1994).
57. Kosslyn, S.M. Image and Mind. Cambridge: Harvard University (1980).
58. Kosslyn, S.M., Ball, T.M., Reiser, B.J. Visual images preserve metric spatial information: Evidence from studies of image scanning. Journal of Experimental Psychology: Human Perception and Performance, 4(1):47–60 (1978).
59. Kosslyn, S.M., Thompson, W.L. When is early visual cortex activated during visual mental imagery? Psychological Bulletin, 129:723–746 (2003).
60. Lehnert, G., Zimmer, H.D. Auditory and visual spatial working memory. Memory & Cognition, 34:1080–1090 (2006).
61. Lehnert, G., Zimmer, H.D. Modality and domain specific components in auditory and visual working memory tasks. Cognitive Processing, 9:53–61 (2008).
62. Logie, R.H. Visuo-Spatial Working Memory. Hove: Lawrence Erlbaum (1995).
63. Logie, R.H. Visuo-spatial processing in working memory. Quarterly Journal of Experimental Psychology, 38A: 229–247 (1986).
64. Logie, R.H., Pearson, D. The inner eye and the inner scribe of visuo-spatial working memory: Evidence from developmental fractionation. European Journal of Cognitive Psychology, 9:241–257 (1997).
65. Logie, R.H., Zucco, G.M., Baddeley, A.D. Interference with visual short-term memory. Acta Psychol (Amst) 75:55–74 (1990).

66. Lovett, M.C., Reder, L.M., Lebiere, C. Modeling Working Memory in a Unified Architecture: An ACT-R Perspective. Cambridge: Cambridge University Press (1999).
67. Luck, S.J., Vogel, E.K. The capacity of visual working memory for features and conjunctions. Nature, 390:279–281 (1997).
68. Mayer, R.E., Sims, V.K. For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. Journal of Educational Psychology, 86:389–401 (1994).
69. Mazard, A., Tzourio-Mazoyer, N., Crivello, F. A PET meta-analysis of object and spatial mental imagery. European Journal of Cognitive Psychology, 16:673–695 (2004).
70. McConnell, J., Quinn, J.G. Complexity factors in visuo-spatial working memory. Memory, 12:338–350 (2004).
71. McConnell, J., Quinn, J.G. Cognitive mechanisms of visual memories and visual images. Imagination, Cognition and Personality, 23:201–207 (2003).
72. McConnell, J., Quinn, J.G. Interference in visual working memory. Quarterly Journal of Experimental Psychology, 53:53–67 (2000).
73. Mecklinger, A., Müller, N. Dissociations in the processing of "what" and "where" information in working memory: An event-related potential analysis. Journal of Cognitive Neuroscience, 8:453–473 (1996).
74. Mellet, E., Bricogne, S., Tzourio-Mazoyer, N., Ghaem, O., Petit, L., Zago, L., Etard, O., Berthoz, A., Mazoyer, B., Denis, M. Neural correlates of topographic mental exploration: The impact of route versus survey perspective learning. Neuroimage, 12:588–600 (2000).
75. Mellet, E., Tzourio, N., Denis, M., Mazoyer, B. Cortical anatomy of mental imagery of concrete nouns based on their dictionary definition. Neuroreport, 9:803–808 (1998).
76. Miller, G.A. The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63:81–97 (1956).
77. Milner, B. Interhemispheric differences in the localization of psychological processes in man. British Medical Bulletin, 27:272–277 (1971).
78. Miyake, A., Friedman, N.P., Rettinger, D.A., Shah, P., Hegarty, M. How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. Journal of Experimental Psychology: General, 130:621–640 (2001).
79. Miyake, A., Shah, P. (Eds.). Models of Working Memory. Mechanisms of Active Maintenance and Executive Control. Cambridge: Cambridge University Press (1999).
80. Oberauer, K. Access to information in working memory: Exploring the focus of attention. Journal Experimental Psychology: Learning, 28:411–421 (2002).
81. Oberauer, K., Süß, H.M., Schulze, R., Wilhelm, O., Wittmann, W.W. Working memory capacity – facets of a cognitive ability construct. Personality and Individual Differences, 29:1017–1045 (2000).
82. Oberauer, K., Süß, H.-M., Wilhelm, O., Wittmann, W.W. The multiple faces of working memory: Storage, processing, supervision, and coordination. Intelligence, 31:167–193 (2003).
83. Owen, A.M. The functional organization of working memory processes within lateral frontal cortex: The contribution of functional neuroimaging. European Journal of Neuroscience, 9:1329–1339 (1997).
84. Pearson, D.G. Imagery and the visuo-spatial sketchpad. In J. Andrade, (Ed.),Working Memory in Perspective (pp. 33–59). Hove: Psychology Press Ltd (2001).
85. Petrides, M. Lateral prefrontal cortex: Architectonic and functional organization. Philosophical Transactions of the Royal Society London B, 360:781–795 (2005).
86. Petrides, M. Dissociable roles of mid-dorsolateral prefrontal and anterior inferotemporal cortex in visual working memory. Journal of Neuroscience, 20:7496–7503 (2000).
87. Postle, B.R., Druzgal, T.J., D'Espositio, M. Seeking the neural substrates of visual working memory storage. Cortex, 39:927–946 (2003).
88. Postle, B.R., Stern, C.E., Rosen, B.R., Corkin, S. An fMRI investigation of cortical contributions to spatial and nonspatial visual working memory. NeuroImage, 11(5 Pt 1):409–423 (2000 May).

89. Pylyshyn, Z.W. Return of the mental image: Are there really pictures in the brain? Trends in Cognitive Science, 7:113–118 (2003).
90. Quinn, J.G. Movement and visual coding: The structure of visuo-spatial working memory. Cognitive Processing, 9:35–43 (2008).
91. Quinn, J.G., McConnell, J. The interval for interference in conscious visual imagery. Memory, 14:241–252 (2006).
92. Quinn, J.G., McConnell, J. Manipulation of interference in the passive visual store. European Journal of Cognitive Psychology, 11:373–389 (1999).
93. Quinn, J.G., McConnell, J. Irrelevant pictures in visual working memory. Quarterly Journal of Experimental Psychology A, 49A:200–215 (1996).
94. Quinn, J.G., Ralston, G.E. Movement and attention in visual working memory. Quarterly Journal of Experimental Psychology, 38A:689–703 (1986).
95. Raffone, A., Wolters, G., Murre, J.M.J. A neurophysiological account of working memory limited capacity: Within-chunk integration and between-item segregation. Behavioral and Brain Science, 24:139–141 (2001).
96. Räamä, P., Poremba, A., Sala, J.B., Yee, L., Malloy, M., Mishkin, M., Courtney, S.M. Dissociable functional cortical topographies for working memory maintenance of voice identity. and location. Cerebral Cortex, 14:768–780
97. Ranganath, C., Cohen, M., Dam, C., D'Esposito, M. Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. Journal of Neuroscience, 24:3917–3925 (2004).
98. Ranganath, C., DeGutis, J., D'Esposito, M. Category-specific modulation of inferior temporal activity during working memory encoding and maintenance. Cognitive Brain Research, 20:37–45 (2004).
99. Redick, T.S., Engle, R.W. Working memory capacity and attention network test performance. Applied Cognitive Psychology, 20:713–721 (2006).
100. Repovs, G., Baddeley, A.D. The multi-component model of working memory: Exploration in experimental cognitive psychology. Neuroscience, 139:5–21 (2006).
101. Rolke, B., Heil, M., Hennighausen, E., Häussler, C., Rösler, F. Topography of brain electrical activity dissociates the sequential order transformation of verbal versus spatial information in humans. Neuroscience Letters, 282:81–84 (2000).
102. Rösler, F., Heil, M., Bajric, J., Pauls, A.C., Hennighausen, E. Patterns of cerebral activation while mental images are rotated and changed in size. Psychophysiologie, 32:135–149 (1995).
103. Ruchkin, D.S., Grafman, J., Cameron, K., Berndt, R.S. Working memory retention systems: A state of activated long-term memory. Behavioral and Brain Sciences, 26:709 (2003).
104. Ruff, C.C., Knauff, M., Fangmeier, T., Spreer, J. Reasoning and working memory: Common and distinct neuronal processes. Neuropsychologia, 41:1241–1253 (2003).
105. Sack, A.T., Hubl, D., Prvulovic, D., Formisano, E., Jandl, M., Zanella, F.E., Maurer, K., Goebel, R., Dierks, T., Linden, D.E.J. The experimental combination of rTMS and fMRI reveals the functional relevance of parietal cortex for visuospatial functions. Brain Research. Cognitive Brain Research, 13:85–93 (2002).
106. Shah, P., Miyake, A. The separability of working memory resources for spatial thinking and language processing: An individual differences approach. Journal of Experimental Psychology: General, 125:4–27 (1996).
107. Shepard, R.N., Metzler, J. Mental rotation of three-dimensional objects. Science, 171: 701–703 (1971).
108. Sims, V.K., Hegarty, M. Mental animation in the visuospatial sketchpad: Evidence from dual-task studies. Memory & Cognition, 25:321–333 (1997).
109. Smith, E.E., Jonides, J. Working memory: A view from neuroimaging. Cognitive Psychology, 33:5–42 (1997).
110. Smyth, M.M., Scholey, K.A. The relationship between articulation time and memory performance in verbal and visuospatial tasks. British Journal of Psychology, 87:179–191 (1996).
111. Song, J.-H., Jiang, Y. Visual working memory for simple and complex features: An fMRI study. Neuroimage, 30:963–972 (2006).

112. Todd, J.J., Marois, R. Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. Cognitive, Affective, & Behavioral Neuroscience, 5: 144–155 (2005).
113. Todd, J.J., Marois, R. Capacity limit of visual short-term memory in human posterior parietal cortex. Nature, 428:751–754 (2004).
114. Treisman, A. Object tokens, binding and visual memory. In H.D. Zimmer, A. Mecklinger, U. Lindenberger (Eds.), Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience (pp. 315–338). Oxford: Oxford University Press (2006).
115. Trojano, L., Grossi, D., Linden, D.E.J., Formisano, E., Hacker, H., Zanella, F.E., Goebel, R., Di Salle, F. Matching two imagined clocks: The functional anatomy of spatial analysis in the absence of visual stimulation. Cerebral Cortex, 10:473–481 (2000).
116. Turner, M.L., Engle, R.W. Is working memory capacity task dependent? Journal of Memory and Language, 28:127–154 (1989).
117. Umla-Runge, K., Zimmer, H.D., Krick, C.M., Reith, W. fMRI correlates of working memory – specific posterior representation sites for movement and position information of a dynamic stimulus (submitted).
118. Ungerleider, L.G., Courtney, S.M., Haxby, J.V. A neural system for human visual working memory. Proceedings of the National Academy of Science USA, 95:883–890 (1998).
119. Ungerleider, L.G., Mishkin, M. Two cortical visual systems. In D.J. Ingle, M.A. Goodale, R.J.W. Mansfield (Eds.), Analysis of Visual Behavior (pp. 549–586.). Cambridge, MA: MIT Press (1982).
120. Vallar, G.E., Shallice, T. (Eds.). Neuropsychological Impairments of Short-term Memory. Cambridge: Cambridge University Press (1990).
121. Ventre-Dominey, J., Bailly, A., Lavenne, F., Lebars, D., Mollion, H., Costes, N., Dominey, P.F. Double dissociation in neural correlates of visual working memory: A PET study. Brain Research: Cognitive Brain Research, 25:747–759 (2005).
122. Verhaeghen, P., Cerella, J., Basak, C. A working memory workout: How to expand the focus of serial attention from one to four items in 10 hours or less. Journal of Experimental Psychology: Learning, 30:1322–1337 (2004).
123. Vogel, E.K.,Woodman, G.F., Luck, J. Storage of features, conjunctions, and objects in visual working memory. Journal of Experimental Psychology: Human Perception Performance, 27:92–114 (2001).
124. Wager, T.D., Smith, E.E. Neuroimaging studies of working memory: A meta-analysis. Cognitive, Affective, & Behavioral Neuroscience, 3:255–274 (2003).
125. Waller, D. Individual differences in spatial learning from computer-simulated environments. Journal of Experimental Psychology: Applied, 6:307–321 (2000).
126. Xu, Y. Encoding color and shape from different parts of an object in visual short-term memory. Perception & Psychophysics, 64:1260–1280 (2002).
127. Zimmer, H.D. Visual and spatial working memory: From boxes to networks. Neuroscience & Biobehavioral Reviews, 32:1373–1395 (2008).
128. Zimmer, H.D. The construction of mental maps based on a fragmentary view of physical maps. Journal of Eductional Psychology, 96:603–610 (2004).
129. Zimmer, H.D. Formkonzepte und Bildmarken: Zwei verschiedene Repräsentationen für visuell-sensorische Merkmale? Sprache und Kognition, 7:40–50 (1988).
130. Zimmer, H.D., Lehnert, G. The spatial mismatch effect is based on global configuration and not on perceptual records within the visual cache. Psychological Research, 70:1–12 (2006).
131. Zimmer, H.D., Münzer-Schrobildgen, M., Troje, N.F. Maintenance of biological movement in working memory (submitted).
132. Zimmer, H.D., Speiser, H.R. The irrelevant picture effect in visuo-spatial working memory: Fact or fiction? Psychologische Beitrage, 44:223–247 (2002).
133. Zimmer, H.D., Speiser, H.R., Seidler, B. Spatio-temporal working-memory and short-term object-location tasks use different memory mechanisms. Acta Psychologica (Amsterdam), 114:41–65 (2003).

From Resource-Adaptive Navigation Assistance to Augmented Cognition

Hubert D. Zimmer, Stefan Münzer, and Jörg Baus

1 Introduction

In an assistance scenario, a computer provides purposive information supporting a human user in an everyday situation. Wayfinding with navigation assistance is a prototypical assistance scenario. The present chapter analyzes the interplay of the resources of the assistance system and the resources of the user. The navigation assistance system provides geographic knowledge, positioning information, route planning, spatial overview information, and route commands at decision points. The user's resources encompass spatial knowledge, spatial abilities and visuo-spatial working memory, orientation strategies, and cultural habit. Flexible adaptations of the assistance system to available resources of the user are described, taking different wayfinding goals, situational constraints, and individual differences into account. Throughout the chapter, the idea is pursued that the available resources of the user should be kept active.

2 Resources

In this section we describe resources that contribute to performance in a navigation task. We start with the user's resources, followed by the resources of the navigation assistance system. Generally, we use "resource" as an umbrella term to refer to entities that are used to reach specific processing or action goals [28]. They are functional for a task and in some sense auxiliary material to fulfill the task demands. Many different entities can be considered as resources. Sensors, effectors, knowledge, and time are resources for all kinds of systems; central processing capacity, storage space, algorithms, and resolution of displays are examples of resources in

H.D. Zimmer (✉)
Brain and Cognition Unit, Department of Psychology, Saarland University, 66123 Saarbrücken, Germany
e-mail: huzimmer@mx.uni-saarland.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_3,

artificial systems; and for biological systems resources are brain structures, neural networks, neurotransmitters, cognitive efficiency, mental processing capacity, etc.

Resources are often limited and their availability determines the efficiency of task solving. The more resources are available or the less effort is necessary for making them available, the higher are the performances of a system. Using resources causes costs, for example, in terms of money, time, space, energy, etc. One therefore can specify utility functions, i.e., the likelihood of success or the quality of the outcome of a process as a function of used resources. Usually, however, the exact shape of this function is not relevant. It is only of interest that the required resources are lower than the available ones because otherwise the system will fail because its resource limitation is exceeded. Therefore in general, one wants to know what resources are available and what their limits are. This knowledge is sufficient to design a task in such a way that it can be successfully processed because it remains below the resource limits. In this chapter we want to discuss some resources that are relevant in navigation and especially when participants use navigation assistance. We want to disclose limitations of these resources and possible side effects that are (unintentionally) induced when navigation systems are used.

2.1 User's Resources

Spatial knowledge: When exploring an environment solely by navigation (i.e., without a map), the majority of people first acquire knowledge about landmarks and routes. Landmark knowledge means to remember appearances of salient places and objects like buildings. Route knowledge can be conceived of associative memory of a sequence of landmarks together with directions. Landmark and route knowledge is based on views from an individual's perspective. This kind of knowledge is anchored in an egocentric reference frame, i.e., it is relative to the individual's position and orientation in the environment. In contrast, survey representations provide an overview over the spatial layout, based on an extrinsic frame of reference, i.e., from an allocentric perspective [18, 23, 39]. Such a mental representation is also called a "mental" or "cognitive" map. It allows flexible spatial orientation, e.g., drawing inferences about spatial relations between places and planning of routes not yet travelled. Gillner and Mallot [21] have shown that people can acquire survey-like knowledge of an environment which is experienced solely by navigation. However, the mental representation is similar to a graph (rather than to a metric map), and it is not necessarily globally coherent. Nevertheless, the quality of the particular mental representation depends on the types of information available when studying the environment. Thorndyke and Hayes-Roth [68] have studied learning based on either the egocentric (extensive navigation experience without map study) or the allocentric perspective (map study without navigation experience). Navigational experience facilitated performances in tasks involving the egocentric perspective (such as route distance estimation and direction estimation) where as map study facilitated performances in tasks involving the allocentric perspective (such as Euclidean distance estimation and locating a destination given two reference points). Thus, depending

on the acquired mental representation, spatial task performances might require additional mental transformations. The orientation specificity of maps is another example. If the orientation of the map is not aligned with the actual orientation of an individual in the real environment, then the representation has to be mentally rotated. This transformation increases errors and computation times [56, 57].

Neuropsychological as well as neuroimaging studies support the distinction between the allocentric and the egocentric reference frame. Both spatial tasks activate a fronto-parietal network; however, there are also specialized regions. For instance, the hippocampal formation seems to play a special role in allocentric spatial processing [46, 48], and partially different networks are involved in allocentric and egocentric transformations [73].

Spatial mental models can also be formed from verbal descriptions of environments [16, 67]. Vice versa, precise verbal descriptions can be formulated from a spatial mental representation after studying a map [67]. However, it has to be noted that there is a superiority of the pictorial, "analogous" presentation format for spatial information. Studying a map resulted in a memory advantage, compared to studying equivalent verbal descriptions [74]. This advantage was also found when map fragments were presented, i.e., when the visuo-spatial information was provided piecemeal as in the verbal descriptions [74]. Similarly, in an unpublished study, Zimmer and collegues could show a picture superiority effect for learning route descriptions. Pictures of landmarks and icons for direction information were compared with proper names and verbal direction information. Memory was best with pictorial information for both landmarks and direction information. Furthermore, the spatial relation of the direction information to the landmark as shown on the display played a role. Memory benefited from spatial positions that supported the represented information (e.g., when an arrow pointing to the half-right was presented on the display 45° to the right of the landmark).

In summary, the acquisition of spatial orientation knowledge and the acquisition of route knowledge appear to profit from visual and "analogous" presentation formats as provided by maps and pictures (and relative positions) of places and landmarks. However, the mental representation of the environment depends strongly on the learning experience with respect to the spatial reference frame involved. Depending on the available representation, effortful mental transformations might be necessary to perform certain spatial tasks, while other tasks are directly supported. In order to optimize the resource "spatial environmental knowledge" we have to take the described properties of the mental representations of the human user into account.

Individual differences: In general, people differ remarkably in their ability to form and to use mental spatial representations of their environment. Individual differences are a typical observation in spatial cognition experiments. This is true both for orientation in real environments (e.g., [38]) and for learning from virtual environments (e.g., [69, 70]).

The individual differences are caused by a number of cognitive and personality factors including mental abilities, spatial strategies, and self-confidence. The variability of these resources has consequences for the interaction of a particular

user with an assistance system. Assistance systems that adapt to the user's resources must consider these components in the user model.

The capacity of working memory: The general function of working memory is to hold information in an accessible state, in order to use the information in ongoing cognitive processing. Working memory is not a monolithic component but a network of separate subcomponents (see Zimmer, Münzer and Umla, Visuo-spatial Working Memory as a Limited Resource of Cognitive Processing). In the present context, however, it is notable that working memory is a central processing resource that varies inter-individually. The individual capacity is related to intelligence and limits performance in everyday higher-order cognitive tasks [13, 17, 29]. Aging is associated with decreases in working memory capacity [20]. This decrease specifically can impair the construction of a coherent spatial mental model [54].

Working memory is considered a key factor in learning about the spatial configuration of a real environment [24]. However, only a few studies have directly investigated the contribution of working memory to environmental learning. For instance, Garden, Cornoldi and Logie [19] demonstrated that secondary tasks had detrimental effects on route learning in a real environment and this interacted with spatial ability of the subjects. High spatial ability participants appeared more affected by a spatial interference task, whereas low spatial ability participants were more affected by a verbal interference task. In a series of experiments involving learning during navigation through a novel, real-world environment, we have found a substantial and robust relation between visuo-spatial working memory capacity and the acquisition of orientation knowledge in a real environment (see below). Similarly, Hegarty et al. [24] have found considerable correlations between a visuo-spatial working memory measure and learning from direct experience in the environment. Thus, the capacity of visuo-spatial working memory appears as a critical resource in real-world spatial orientation tasks.

Spatial abilities: Spatial abilities are traditionally measured by paper-and-pencil tests (e.g., as part of intelligence testing). Such tests require mentally visualizing some manipulation of an object in space, for example, mentally rotating a three-dimensional object or mentally folding a piece of paper. Surprisingly, it has repeatedly been found that spatial ability tests are only weakly related to environmental spatial performance (e.g., [3], see [26], for a review). It has therefore been suggested that spatial abilities as measured by those tests should be distinguished from environmental spatial abilities [26]. The latter, but not the former require integrating different views of the environment over time to form a coherent mental representation. Spatial ability tests, in contrast, require the comparison and/or mental manipulation of single objects that can be apprehended in a single view. Utilizing a latent variable approach, it was found that the latent variable representing spatial abilities was substantially related to learning from a real environment [24]. The latent variable, however, also included a visuo-spatial working memory measure. In contrast to learning in the real environment, spatial abilities do predict visuo-spatial learning from visual media [24], e.g., by inspecting virtual environments. Waller [69] has found that spatial abilities explain a part of the remarkable individual differences in spatial learning with virtual environments.

In summary, spatial abilities as traditionally measured with paper-and-pencil tests do not appear to play a critical role in spatial learning in the real environment; however, spatial abilities are an important resource when learning spatial configurations from visual media.

Environmental spatial strategies: Since spatial tasks can often be solved by either allocentric or egocentric processing, individuals have a choice. This results in strategic individual differences. Aginsky et al. [1] described two different strategies to solve a route retrieval task in a driving simulator. The first strategy relied more on the appearances of landmarks at intersections and was therefore termed the "visually dominated strategy." The second strategy relied more on the knowledge of the relations between landmarks and was therefore termed the "spatially dominated strategy." Participants with either the visually or the spatially dominated strategy differed with respect to the types of errors they made. The visual vs. spatial strategies correspond to the egocentric vs. the allocentric reference frame.

In a recently developed questionnaire on spatial representation [15, 50], individuals are asked about their ways to code and represent large-scale space. The distinctions between landmark-centered, route-based, and survey-like representations are understood as preferred and stable strategies of individuals, which might predict actual formation of a mental representation of the environment [51]. The distinction between these two strategies is also supported by functional MRI. Different neural structures were active dependent on the type of strategy participants indicated [8] and dependent on their spatial abilities [61]. Available strategies thus are an important resource of an individual to orient in large-scale space.

Self-reported competency: The everyday concept "sense of direction" reflects self-estimated or ascribed individual competence differences in orientation ability. Research on the sense of direction suggests that people can give reliable and valid estimates of their environmental spatial abilities, which correlate substantially with actual performance in the environment [33, 24, 25]. Recently, a (unitary) 15-item scale has been developed (Santa Barbara Sense of Direction Scale SBSOD, [25]). The scale includes questions on using maps, giving and understanding directions, remembering routes, doing the navigational planning, and utilizing cardinal directions. In the Hegarty et al. study [24], this scale explained a considerable portion of the variance in spatial layout learning in a real-world task. However, the measure might capture also a kind of self-confidence rather than a competence, and the underlying everyday concept appears to be culture-specific (see next section).

*Cultural habits and the man-made environment: S*patial processing depends on cultural habits [37]. For instance, Westerners commonly have a more analytic and object-centered way of perception whereas Asians tend to take the context more into account [47]. Cultural influences can therefore foster a rather "analytic" vs. a more "holistic" spatial processing of scenes [41].

Moreover, the man-made environment and the common way to use it for wayfinding might play an important role. In the US the dominant man-made environmental element seems to be the straight line. Layouts of city streets are often grid-like. Names of streets as well as directional signing on motorways refer to the cardinal

directions. This way of providing spatial information suggests using an allocentric references frame (like "going north") and facilitates perspective transformation. In contrast, European environments do not provide allocentric information in street layouts and signing. These differences might explain why the SBSOD (developed in the US) is a unitary scale although it involves both allocentric and egocentric reference frames, while the QSR (developed in Europe) distinguishes between strategies in different reference frames.

Gender and age: Gender is often thought to be a critical factor in spatial abilities, with men outperforming women in particular spatial tasks such as mental rotation [36]. Orientation in the real environment, however, does not show such a clear superiority. In their review on gender differences in orientation performance, Coluccia and Louse [12] found that gender differences depended on learning conditions and kind of test. Differences favoring men were found in those studies in which orientation tasks were particularly difficult. Coluccia and Louse concluded that individual differences in orientation performance may reflect individual differences in visuospatial working memory resources. Following this conclusion, gender differences appear to reflect processing capacities rather than specific performance differences.

The same is true for age. Age concerns the capacity of working memory; thus older people will perform similarly to younger people who have comparable working memory capacity. For example, when driving performance for elderly drivers was compared to that of younger participants, attention and visuo-spatial abilities were the best predictors independent of the age of the drivers [4]. A comparable result was observed by Zimmer and collegues in an unpublished study. They compared route learning from verbal descriptions with route learning from pictures and icons, for older people (older than 65) and students. It was suggested that older participants show a specific performance decrement in visuo-spatial working memory tasks [29]. As a consequence, verbal descriptions in route learning might be more appropriate for elderly people than for younger people. However, both groups showed a clear pictorial advantage, and younger participants performed better than older participants as was observed in working memory tasks. This finding supports the superiority of the visual modality for route learning independent of age and the hypothesis of general impairment in older age.

The bottleneck of central attention and cognitive control: Many everyday situations require the coordination of two activities at the same time. For instance, while driving a car, people listen to the radio, attend to the multimedia entertainment, or follow the navigation assistance system. This situation has been analyzed as a dual-task situation. While driving in heavy traffic or being on the phone under bad reception conditions, performance decrements result due to interference of the concurrent tasks (e.g., [9, 10, 53]).

Research on dual-task coordination has shown that reaction times and performance errors increase if a reaction is required to a secondary task before the response preparation and production of the reaction to a primary task could be completed. This has mainly been interpreted as caused by capacity limitations at the response selection stage [49]. This stage is limited to process one event at a time. However, also the encoding of an event in the secondary task might be impeded if

processing the primary task is still ongoing [45]. Thus it can be expected that under demanding conditions in which new events in a primary task (e.g., driving) and in a secondary task (e.g., manipulating a car interface) appear in close succession, the cognitive system is overstrained regardless of the modality of the incoming information. For example, it has been shown that hands-free cell phone conversations impeded driver's reactions to vehicles breaking in front of them, because of reduced attention to visual input [65].

For the design of assistance systems to be used in cars it is thus important to realize the existence of a bottleneck that concerns central attention, irrespective of modality. The competition for the visual modality is an additional constraint [59]. Again, the ability to control attention is subject to individual differences. One important diagnostic variable is age. Age increases the susceptibility to interference in dual-task situations [55].

2.2 System's Resources

Nowadays, navigation systems come in a variety of forms. Some are running as an internet service, accessed via web-browser, which allow a user to calculate specific route information prior to a trip; others are used in automotive navigation. Almost every automobile manufacturer offers a proprietary type and additionally, mobile navigation systems of independent manufacturers exist running on especially designed portable small devices offering complete car/bike navigation solutions. Those systems are accompanied by systems designed to run on Personal Digital Assistants and mobile phones supposed to lead you whether you travel by car, by bike or on foot. And last but not least, a vast amount of research prototypes were especially designed for pedestrian navigation (for an overview see [5]). Nearly all of those systems, at least the commercially available ones, provide incremental navigation instructions by means of spoken and visual instructions and they all rely on various technical resources. Technical resources cover all types of limitations of the presentation platform. Here, four different subtypes can be identified:

- Quality of the positioning information
- Quality of the geographical database
- Quality of the presentation
- Computational restrictions

Quality of Positioning Information: Navigation systems designed for use in automotive navigation typically use the Global Positioning System (GPS) to locate vehicles in the outdoor environment. Due to the fact that urban canyons and tunnels may decrease the quality of GPS signals or cause a complete signal loss, automotive navigation systems use different additional sensors to improve positioning information and reliability. Information about the travel speed, the route, and the distance travelled so far is used in dead reckoning and map matching algorithms to compute the vehicles' position on the digital road network provided by the database. Pedestrian

navigation systems in an outdoor scenario have to cope with the same problem. However, these systems typically cannot rely on additional sensors to improve positioning information. Furthermore, since several different means of transportation typically have to be combined in order to reach a destination, it must be ensured that the navigation system reacts to the user's changing situation, regardless of the kind of transportation. When a change of the means of transportation is detected, the system should adapt to the new situational constraints. The essential switch between different positioning technologies, e.g., GPS or GSM/UTMS cells (outdoors) or infrared, ultrasound, Bluetooth, W-LAN, and RFID-Tags (indoors), should ideally remain unnoticed by the user such as firstly described in [6].

Quality of the database: Digital maps for automotive navigation systems are commercially available and represent information about the street network; in addition to paper based street maps, they offer various points of interests, e.g., gas stations, railway stations, or public buildings. Nowadays, several systems are available, e.g., TomTom™, that allow their users to annotate and share their maps in order to keep them up to date. In the case of pedestrian navigation systems the situation differs from automotive navigation, since the pedestrian users are not bound to follow paths or streets. Instead they typically cross open spaces, directly following the line of sight; therefore the geographic database has to reflect this and represent places as polygonal objects, in contrast to commercial street map databases which usually consist of line segments. Furthermore, in order to allow for pedestrian navigation inside of buildings the navigation systems' database has to provide models for complex, multi-level buildings too. A system especially designed to allow for pedestrian map modeling has been developed by [64]. As for automotive maps, such digital databases have to be updated on a regular basis to reflect the changes in the environment.

Quality of the presentation: As mentioned before navigation systems use a combination of speech and graphics to present incremental navigation instructions. The more sophisticated such presentations or the more technical resources are required to generate them. Concerning speech, the resources to generate simple verbal instructions, such as "Turn right after 500 m," can be kept relatively low. Concerning the visualization of navigational instruction mostly all systems comprise a combination of different visualization techniques including 2D and/or 3D visualizations of route segments, with the possibility to align the map to the user's current view, which means that the visualized map has to be constantly rotated, leading to increased computational resources but reducing mentally costly and error-prone perspective transformations. In addition to the visualization of maps, many systems use icons to visualize turning actions and display numbers as quantitative information, e.g., to show the next navigational action and the distance to the next turning point. Generally, the amount of computational resources needed to present navigational instructions increases with the amount of details that has to be displayed.

Computational restrictions: The quality of navigational instructions generated by navigation systems depends on different technical resources, namely the quality of positioning information, quality of the digital road database, quality of sensor data

and algorithms used to improve positioning information, and the size and resolution of the display used to visualize navigation instructions. These resources can be subsumed as computational resources of the navigation system, which are restricted. Especially, in a mobile scenario these restricted resources impact the systems performance, since mobile devices are more severely limited in terms of computational power, than in-car or stationary systems. Nevertheless, technical advances in hardware development over the last years and the ongoing tendency that computational power will further increase in the future led and will lead to sophisticated navigation systems, where computational resources will play a minor role.

3 Goals

After having described the resources, we will now turn to their interaction in the accomplishment of goals in wayfinding situations. A primary goal is finding the way to the destination. An important secondary goal is the incidental spatial learning during navigation. Even the ubiquitous availability of a navigation system would not make user's spatial knowledge superfluous. It is still used in planning routes, in autobiographical memory, and in spatial reasoning – and it is a resource that can be utilized by an intelligent assistance system. In addition, the consequences of possible malfunctions of the assistance system provide pragmatic reasons for learning. If the assistance system fails, people will experience severe difficulties to orient in the unknown environment. We have therefore investigated the attainment of this secondary goal during assisted navigation.

The studies reported below consider two ecologically valid scenarios for pedestrian navigation. In the first scenario, a visitor takes a guided tour through a novel environment. Spatial knowledge is acquired incidentally, i.e., as a by-product of the navigation and exploration activity. In the second scenario, a visitor needs to find a particular office in a complex building. Here, the wayfinding goal comes to the fore, since the visitor is not interested in exploration, and learning of the individual route is intentional in this scenario.

Incidental learning from navigation assistance when exploring a novel environment: A side-effect of using a traditional map in a wayfinding task is the incidental learning about the spatial configuration of the environment. In contrast, much less is learned about the spatial configuration if navigation assistance is used [44]. We have suggested that map usage requires active encoding of the spatial information for the purpose of wayfinding (involving perspective transformation, mental rotation, and rehearsal of route instructions) and that this active encoding supports spatial learning. The comfort of navigation assistance thus impedes learning by letting people follow the route passively. Presenting allocentric information on the assistance system in addition to route commands was not sufficient to activate learning processes. The idea of active encoding is derived from the transfer-appropriate-processing principle [42]. Optimal performances are achieved when the information processed during encoding matches the information needed during testing.

In a second series of experiments we applied this principle to wayfinding assistance on a PDA. Wayfinding comfort was slightly reduced, but it was expected that these costs would be compensated by incidental learning of the spatial layout. In all experiments, first-time visitors to the campus of Saarland University took a guided tour passing salient places, using navigation assistance. After the tour, participants were asked (1) to complete a route recognition test, (2) to estimate directions between visited places (pointing), and (3) to draw a sketch map. Across experiments, the wayfinding presentation on the navigation assistance varied with respect to modality (verbal vs. visual), perspective (egocentric vs. allocentric), alignment (north-aligned vs. rotated), and completeness of information (map fragments vs. complete map; street-layout-based route commands vs. directional information without street layout). Route knowledge was generally not affected by presentation variations. Presumably, route knowledge is acquired from direct experience in the environment, and redundant ground-plane, egocentric information on the assistance system does not improve this kind of orientation knowledge. However, better survey knowledge was acquired with all presentation formats that provided allocentric spatial information. Allocentric information included north-aligned and rotated map fragments, north-aligned complete maps, and directional information without information about the street layout. The specific format of the allocentric information thus did not affect survey knowledge; however, it might be noted that the encoding of the information presented might have required inference processes of different quality (i.e., perspective transformation with a map-based presentation vs. integration of information from the environment when the presented information was incomplete). It can be concluded that the way in which information is presented on assistance systems affects the mental representation of the environment, although the information is acquired incidentally as a by-product of wayfinding. This is caused by mental processes that encode the presented information, and compare and integrate it with the information in the environment. With allocentric information, this processing is more active (involving some inferences) and therefore (according to transfer-appropriate processing), it allows more flexible access from long-term memory in subsequent tests.

The experiments additionally contributed to an explanation of individual differences. A measure of visuo-spatial working memory capacity was obtained from every participant. We observed in all of the experiments that the angular error resulting from the direction estimation tasks (indicating environmental knowledge) was substantially related to visuo-spatial working memory capacity, with the higher the visuo-spatial working memory capacity, the lower the angular error. It can therefore be concluded that the formation of a mental survey representation from direct experience is limited by individual visuo-spatial working memory resources.

Intentional learning of route descriptions in a complex building: Individual short-time visitors to large, complex buildings such as conference centers might be provided with personalized route descriptions that guide them to a particular destination. The scenario differs from the preceding one in that a temporary route instruction should be learned as efficient as possible for immediate use. In the scenario, users are presented with individual visual route descriptions on ambient

displays in the environment. Going without mobile navigation pragmatically circumvents the technical indoor positioning difficulties. The route descriptions are provided by the modeling software Yamamoto [64] which calculates individual paths through a virtual model of the building. The question of the study concerned the instructional format. A movie of the route through a virtual model of the building taken from the egocentric perspective and a sequence of pictures showing the succession of decision points with route indicators (arrows), also seen from the egocentric perspective, were compared to a route depicted on a ground plan (allocentric perspective). After getting the instruction, participants had to find their way through the building on their own. Results showed that the movie-based route instruction was most appropriate. Although the movie included the highest amount of information and did not show overview information, it probably provided the best cues for navigation because it directly matched the visual experience during walking. This is another example of the transfer-appropriate-processing principle.

Driving and being guided: As a driver, the user has the goal to be guided reliably and comfortably to the destination. Navigation assistance systems of today are built primarily to accomplish this goal. Particularly, they provide route instructions auditory-verbally and spatial overview information visually, which appears appropriate because providing voice directions causes less driving errors than map guidance [66].

However, there are some challenges in the automotive situation. The first challenge is to present the user with route instructions that can easily be understood. The second challenge is to provide helpful spatial overview and orientation information for the user in addition to route commands. The third challenge is to account for the dual-task demands during driving. Finally, the user should keep aware of the wayfinding activity such that he/she does not blindly follow the route commands.

Navigation assistance systems of today do not provide satisfactory solutions to these challenges. While driving in a city center, the route instructions are often difficult to follow. This is because the commands comprise metric distances instead of sequential, qualitative information and refer to street names and route numbers instead of salient landmarks and signing. The screen size of the navigation assistance system restricts the space for spatial overview information in an appropriate resolution. Detailed discussions of these aspects are found in [2, 11, 30, 63]. Furthermore, analyzing the visual display and watching the traffic while driving increases the dual-task demands for the user [27]. However, also voice guidance can increase cognitive load if the instructions are difficult to comprehend [22]. Finally, navigation assistance systems are not inerrable. Erroneous route instructions remain unnoticed if the driver does not take global orientation knowledge into account. Not only have people found themselves far away from their destinations, but also severe accidents have happened because of mindless adherence to navigation assistance.

Current technical developments and cognitive research contribute to improvements of navigation assistance in the automotive situation. These developments are directly targeted on the reduction of cognitive overload. For instance, head-up displays which show visual route command indicators (visual arrows) in the visual field of the driver serve to decrease visual distraction because attention to the two

tasks is oriented to the same spatial location [62]. Furthermore, there is research on cross-modal spatial attention. If the content of a route command is congruent with the direction from which it is presented (i.e., the command "to the left" comes from the left speaker), then processing the spatial information might need less attention than when the information is presented from the center or from the opposite direction [32, 60]. Furthermore, sensor systems in the car can provide input to navigation systems as investigated by [43]. Such in-car assistance systems can use the car's serial-production sensory equipment, e.g., distance control, acceleration sensor, etc. to estimate the driver's current cognitive load. In case of cognitive overload information that is not time-critical can be delayed until the driver's cognitive load decreases. Finally, there are important achievements in verbal-auditory man–machine communication which bring forward the verbal channel for command and information, thereby relaxing the need for visual information in the car [58].

These developments are suitable to avoid cognitive load, stress, and errors in wayfinding. This probably is of particular relevance for older users. Typical causes for accidents with older drivers indicate cognitive overload, e.g., driving errors at intersections and junctions, rear-end collisions, and overlooked traffic signs [14, 52]. It can be expected that older drivers show slower reactions in critical dual-task situations such as following route commands of a navigation assistance system in a novel city center [40]. The above mentioned developments might help to reduce the load. However, more specific developments are desirable for resource adaptation in navigation assistance especially if secondary goals are considered, as for example acquisition of spatial knowledge or allowing for driver's personal preferences. These include route planning algorithms that take simplicity and learnability into account, cognitively adequate survey information, and a user model that accounts for individual differences (see next section).

4 Assistance Systems of the Future: Augmented Cognition

The authors believe that navigation assistance systems of the future might be intelligent cognitive aids that support users in a wide range of wayfinding scenarios. This encompasses cognitively appropriate and individually tailored route instructions as well as support for learning about the environment. Furthermore, the assistance system might provide an external autobiographical memory for the user. Communication with the user might be based on a model of the user and the situation. This model includes the user's knowledge and preferences, his/her spatial strategies, and his/her current mind states (e.g., motivational and may be even emotional state in the current situation). Assistance systems of the future thus adapt to and supplement the user's individual resources (knowledge, mental abilities, strategies, and attentional resources).

In the following, we will describe possibilities for adaptation to the available resources primarily on the side of the system. However, also users might adapt and utilize their resources, particularly concerning the acquisition and usage of global

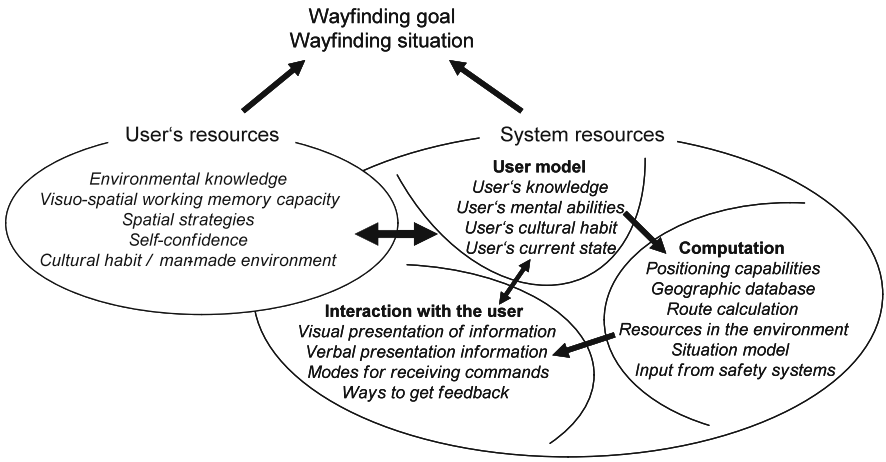


Fig. 1 Resources of the user and system interact to achieve a common goal

spatial knowledge and spatial strategies. In general, the resources of the user and the resources of the system interact and complement each other (see Fig. 1).

User model: In the user model, the navigation system keeps long-lasting diagnostic information about the user's resources with respect to knowledge, mental abilities and strategies, as well as cultural habit. The user model is regularly updated. Some information can be made accessible by storing past experiences. For instance, knowledge can be estimated by storing all travelled routes; strategies can be estimated by past preferences for route calculation and information presentation. Attentional resources can partially be inferred by the age of the user or performances in previous tasks. Some more diagnostic information like sense of direction or self-reported global spatial knowledge could be asked for. A user who is willing to "cooperate" with the assistance system might accept giving this information which is needed for the system's adaptation. Even tests on spatial orientation (offered as games) and specific learning modes in which the system presents additional spatial information about the environment could be included. All this would result in a quite detailed user model and consequently in a system that adapts accordingly to the user model, resulting in a personalized navigation system for the particular user.

Situation model: In contrast to the long-lasting user model, the situation model represents the short-term factors that influence the adaptation of the assistance system. The situation model incorporates the current wayfinding scenario and goal, as well as the current settings for route calculation and information presentation (which are also influenced by the user model). Furthermore, the situation model includes information about the current state of the user which might modify the long-lasting preferences (e.g., motivation, current interest, and stress). In order to accommodate for the dual-task situation, the situation model needs to gather information about the world outside (i.e., density of traffic, driving situation, and possibly dangerous situations) and it needs to be rapidly updated concerning this information as envisioned in [71, 72].

Geographic database: The most important factor for easy understanding of route commands is the alignment between the verbal names or visual icons of landmarks (i.e., signs, places, buildings, streets, bridges, etc.) denoted by the assistance system and the actual identification of these landmarks in the real environment. The reference to street names and route numbers, which are difficult to detect in the real environment, should be replaced. Referring to the signing and to "salient" landmarks makes identification easier. Admittedly, it is costly to enrich geographic databases with these types of information. The determination what a "salient" landmark requires either human evaluation or the possibility to define and add landmarks to the digital database, which are salient to the current user (see also [7, 35]). A resource that is implicitly utilized when creating such detailed databases are cultural differences regarding the spatial information provided in the man-made environment. For instance, signing on motorways in Germany shows directions to distant cities, while signing on motorways in the US (as well as street names in many cities) refer to cardinal directions. If different navigation systems could communicate in a network, their shared knowledge could be used, e.g., where do clusters of wayfinding errors occur, where is cognitive load usually increased, etc.

Positioning and orientation: Positioning and orientation are seldom a problem in the automotive navigation situation. However, in pedestrian navigation scenarios the orientation information might be absent, or both positioning and orientation are not available (e.g., indoor navigation). It is possible, however, to gather this information on alternative ways, e.g., by utilizing the resources of the user (his/her eyes and his/her environmental knowledge). The prerequisites are appropriate modes of interaction. For instance, if the navigation system in the pedestrian situation knows the position, but not the orientation, it might generate a series of egocentric 3D-views of the near surroundings from the geographic database. It might then ask the user which picture comes closest to the current view at his/her position as introduced in Kray [34].

Route calculation: Route calculation should be flexible. Current algorithms find the shortest (or the least time-consuming, or, the most economic) path. However, the resulting route is often difficult to understand and, consequently, difficult to learn. An arguable longer, but much simpler route can help to avoid errors and stress during driving as mentioned in Baus et al. [5]. However, there are a number of ways to define what a simple route is. A simple route may use main streets, contain easily identifiable objects along the way (e.g., salient landmarks and buildings, good signing), and/or refers to places that the user already knows. Furthermore, a simple route is easy to understand with respect to global orientation knowledge (e.g., "we will go always parallel to the river") or to signing ("in the first half of the route we will follow the airport sign").

Information presentation and communication: The information presentation module concludes how the user should be provided with route and overview information based on the user and the situation models. As an example, future navigation systems should allow for a greater flexibility in the presentation of auditory navigation instructions. They should allow their users to fine tune the instructions given by the system, such as timing conditions for verbal instructions at decision

points, the structure of verbal instructions, e.g., whether they prefer quantitative information presentation, or references to landmarks related to the forthcoming navigation action. In addition, in car navigation systems it would also be possible to use the stereo capabilities of the car's entertainment system to render verbal instructions in such a way that they appear spatially congruent to the next turning action.

Stretching to the future – From navigation assistance to a personal companion: The future navigation assistance system in the car might act as an intelligent co-driver. That means that the systems might act more and more like a human co-driver, particularly in knowing about the constraints of the human user when adapting to the requirements of the situation. For instance, the intelligent co-driver observes the traffic situation and stops talking if a traffic situation is evaluated dangerous in relation to the drivers' skills and attentional resources. The intelligent co-driver selects landmarks for a route instruction that the driver can see. Also, the intelligent co-driver provides global orientation information verbally, taking the user's knowledge into account ("we will go direction Berlin during the first half of the route" and "we will cross the river and then go into the direction of the harbor"). Furthermore, the intelligent co-driver can understand and answer feedback questions asked by the driver, such as "to the left, at the church?" This kind of assistance allows for correction of the situation model by means of verbal communication.

The pedestrian navigation system of the future, on the other hand, might act as a personal guide that not only offers a range of navigation and tourist services. On request, the assistance system might train the user on spatial knowledge and orientation skills. For instance, while sitting in a cafe, the user (as a tourist in a city center) can play back the recorded route which has just been traveled. The route for the afternoon will be planned according to suggestions made by the system and the selections by the user. Some important places and their spatial relations might be presented by the system and learned in advance by the user. During travel, the assistance system adapts subsequent information presentation to the already existing knowledge of the user. By storing all traveled routes, the system might serve as an extension of the autobiographical memory of the user. In summary, the assistance system of the future might be more than a useful device for navigation; it might behave like a personal companion.

References

1. Aginsky, V., Harris, C., Rensink, R., Beusmans, J. Two strategies for learning a route in a driving simulator. Journal of Environmental Psychology, 17:317–331(1997).
2. Akamatsu, M., Yoshioka, M., Imacho, N., Daimon, T., Kawashima, H. Analysis of driving a car with a navigation system in an urban area. In Y.I. Noy (Ed.), Ergonomics and Safety of Intelligent Driver Interfaces (pp. 85–96). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc (1997).
3. Allen, G.L., Kirasic, K.C., Dobson, S.H., Long, R.G. Predicting environmental learning from spatial abilities: An indirect route. Intelligence, 22:327–355 (1996).

4. Baldock, M.R.J., Mathias, J.L., McLean, J., Berndt, A.: Self-regulation of driving and older drivers' functional abilities. Clinical Gerontology, 30:53–70 (2006).
5. Baus, J., Krüger, A., Stahl, C. Resource-adaptive personal navigation. In O. Stock, M. Zancanaro (Eds.), Mulitmodal Intelligent Information Presentation (pp. 71–93). Berlin: Springer (2005).
6. Baus, J., Krüger, A., Wahlster, W. A resource-adaptive mobile navigation system. IUI 02 – International Conference on Intelligent User Inter-faces (pp. 15–22). ACM Press, New York (2002).
7. Baus, J., Wasinger, R., Aslan, I., Krüger, A., Maier, A.M., Schwartz, T. Auditory perceptible landmarks in mobile navigation. IUI 07 – 2007 International Conference on Intelligent User Inter-faces (pp. 302–304). ACM Press, New York (2007).
8. Bohbot, V.D., Iaria, G., Petrides, M. Hippocampal function and spatial memory: Evidence from functional neuroimaging in healthy participants and performance of patients with medial temporal lobe resections. Neuropsychology, 18:418–425 (2004).
9. Briem, V., Hedman, L.R. Behavioural effects of mobile telephone use during simulated driving. Ergonomics, 38:2536–2562 (1995).
10. Brookhuis, K.A., de Vries, G., de Waard, D. The effects of mobile telephoning on driving performance. Accident Analysis and Prevention, 23:309–316 (1991).
11. Burnett, G., Joyner, S. An assessment of moving map and symbol-based route guidance systems. In Y.I. Noy (Ed.), Ergonomics and Safety of Intelligent Driver Interfaces (pp. 115–137). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc (1997).
12. Coluccia, E., Louse, G. Gender differences in spatial orientation: A review. Journal of Environmental Psychology, 24:329–340 (2004).
13. Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., Engle, R.W. Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin & Review, 12:769–786 (2005).
14. Cox, A.B., Cox, D.J. Compensatory driving strategy of older people may increase driving risk. Journal of American Geriatrics Society, 46:1058–1059 (1998).
15. Denis, M., Pazzaglia, F., Cornoldi, C., Bertolo, L. Spatial discourse and navigation: An analysis of route directions in the city of Venice. Cognitive Psychology, 13:145–174 (1999).
16. Denis, M., Zimmer, H.D. Analog properties of cognitive maps constructed from verbal descriptions. Psychological Research, 54:286–298 (1992).
17. Engle, R.W., Kane, M.J., Tuholski, S.W. Individual differences in working memory capacity an what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In P.S. Akira Miyake (Ed.), Models of Working Memory (pp. 102–134). Cambridge: Cambridge University Press (1999).
18. Evans, G.W., Pezdek, K. Cognitive mapping: Knowledge of real-world distance and location information. Journal of Experimental Psychology: Human Learning, 6:13–24 (1980).
19. Garden, S., Cornoldi, C., Logie, R.H. Visuo-spatial working memory in navigation. Applied Cognitive Psychology, 16:35–50 (2002).
20. Gazzaley, A., Sheridan, M.A., Cooney, J.W., D'Esposito, M. Age-related deficits in component processes of working memory. Neuropsychology, 21:532–539 (2007).
21. Gillner, S., Mallot, H.A. Navigation and acquisition of spatial knowledge in a virtual maze. Journal of Cognitive Neuroscience, 10:445–463 (1998).
22. Harbluk, J.L., Noy, Y.I., Trbovich, P.L., Eizenman, M. An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. Accident Analysis and Prevention, 39:372–379 (2007).
23. Hart, R.A., Moore, G.T., Downs, R.M., Stea, D. The development of spatial cognition: A review. In Image and Environment: Cognitive Mapping and Spatial Behavior (pp. 246–288). New Brunswick, NJ, USA: AldineTransaction (1973).
24. Hegarty, M., Montello, D.R., Richardson, A.E., Ishikawa, T., Lovelace, K. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. Intelligence, 34:151–176 (2006).

25. Hegarty, M., Richardson, A.E., Montello, D.R., Lovelace, K., Subbiah, I. Development of a self-report measure of environmental spatial ability. Intelligence, 30:425–448 (2002).
26. Hegarty, M.,Waller, D.A. (Eds.). Individual Differences in Spatial Abilities. New York, NY, USA: Cambridge University Press (2005).
27. Horberry, T., Anderson, J., Regan, M.A., Triggs, T.J., Brown, J. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. Accident Analysis & Prevention 38(1):185–191 (2006).
28. Jameson, A., Buchholz, K. Einleitung zum Themenheft "Ressourcenadaptive kognitive Prozesse". Kognitionswissenschaft, 7:95–100 (1998).
29. Jenkins, L., Myerson, J., Joerding, J.A., Hale, S. Converging evidence that visuospatial cognition is more age-sensitive than verbal cognition. Psychology of Aging, 15:157–175 (2000).
30. Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., Engle, R.W. The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology: General, 133:189–217 (2004).
31. Kimura, K., Marunaka, K., Sugiura, S. Human factors considerations for automotive navigation systems – Legibility, comprehension, and voice guidance. In Y.I. Noy (Ed.), Ergonomics and Safety of Intelligent Driver Interfaces (pp. 153–167). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc (1997).
32. Klatzky, R.L., Marston, J.R., Giudice, N.A., Golledge, R.G., Loomis, J.M. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. Journal of Experimental Psychology: Applied, 12:223–232 (2006).
33. Kozlowsky, L.T., Bryant, K.J. Sense of direction, spatial orientation, and cognitive maps. Journal of Experimental Psychology: Human Perception Perform, 3(4):590–598 (1977).
34. Kray, C. Situated Interaction on Spatial Topics (vol. 274). Berlin: Aka Verlag (2003).
35. Krüger, A., Butz, A., Müller, C., Stahl, C., Wasinger, R., Steinberg, K.-E., Dirsch, A. The connected user interface: realizing a personal situated navigation service. IUI 04 – 2004 International Conference on Intelligent User Inter-faces (pp. 161–168). ACM Press, New York (2004).
36. Linn, M.C., Peterson, A.C. Emergence and characterization of sex-differences in spatial ability: A meta-analysis. Child Development, 56:1479–1498 (1985).
37. Majid, A., Bowerman, M., Kita, S., Haun, D.B.M., Levinson, S.C. Can language restructure cognition? The case for space. Trends in Cognitive Science, 8:108–114 (2004).
38. Malinowski, J.C., Gillespie,W.T. Individual differences in performance on a large-scale, real-world wayfinding task. Journal of Environmental Psychology, 21:73–82 (2001).
39. McNamara, T.P., Ratcliff, R., McKoon, G. The mental representation of knowledge acquired from maps. Journal of Experimental Psychology: Learning, 10:723–732 (1984).
40. Merat, N., Anttila, V., Luoma, J. Comparing the driving performance of average and older drivers: The effect of surrogate in-vehicle information systems. Transport Research F, 8:147–166 (2005).
41. Miyamoto, Y., Nisbett, R.E., Masuda, T. Culture and the physical environment. Holistic versus analytic perceptual affordances. Psychological Science, 17:113–119 (2006).
42. Morris, C.D., Bransford, J.D., Franks, J.J. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16:519–533 (1977).
43. Müller, J. Beanspruchungsschätzung im Automobil mit Bayes'schen Net-zen. Universität des Saarlandes Under non-disclosure agreement until 05/30/2008 (2005).
44. Münzer, S., Zimmer, H.D., Schwalm, M., Baus, J., Aslan, I. Navigation assistance and the acquisition of route and survey knowledge. Journal of Environmental Psychology, 26:300–308 (2007).
45. Müsseler, J., Steininger, S., Wühr, P. Can actions affect perceptual processing? Quarterly Journal of Experimental Psychology A, 54:137–154 (2001).
46. Nadel, L., Hardt, O. The spatial brain. Neuropsychology, 18:473–476 (2004).

47. Nisbett, R.E., Miyamoto, Y. The influence of culture: Holistic versus analytic perception. Trends in Cognitive Science, 9:467–473 (2005).
48. O'Keefe, J., Nadel, L. The Hippocampus as a Cognitive Map. Oxford: Oxford University Press (1978).
49. Pashler, H. Comment on McLeod and Hume: Overlapping mental operations in serial performance with preview: Typing. Quarterly Journal of Experimental Psychology A, 47:201–205 (1994).
50. Pazzaglia, F., Cornoldi, C., De Beni, R. Differenze individuali nella rappresentazione dello spazio: Presentazione di un questionario autovalutativo [Individual differences in spatial representation: A self-rating questionnaire]. Giornale Italiano di Psicologia, 3:241–264 (2000).
51. Pazzaglia, F., De Beni, R. Strategies of processing spatial information in survey and landmark-centred individuals. European Journal of Cognitive Psychology 13:493–508 (2001).
52. Praxenthaler, H. [Automobile driving by elderly persons and traffic safety]. FortschrMed, 111:249–251 (1993).
53. Radeborg, K., Briem, V., Hedman, L.R. The effect of concurrent task difficulty on working memory during simulated driving. Ergonomics, 42:767–777 (1999).
54. Radvansky, G.A., Copeland, D.E. Memory retrieval and interference: Working memory issues. Journal of Memory and Language, 55:33–46 (2006).
55. Riby, L., Perfect, T., Stollery, B. The effects of age and task domain on dual task performance: A meta-analysis. European Journal of Cognitive Psychology, 16:863–891 (2004).
56. Rossano, M.J., Moak, J. Spatial representations acquired from computer models: Cognitive load, orientation specificity and the acquisition of survey knowledge. British Journal of Psychology, 89:481 (1998).
57. Rossano, M.J.,Warren, D.H. Misalignedmaps lead to predictable errors. Perception, 18: 215–229 (1989).
58. Salvucci, D.D., Macuga, K.L., Gray, W., Schunn, C. Predicting the effects of cellular-phone dialing on driver performance. Cognitive Systems Research, 3:95–102 (2002).
59. Santos, J., Merat, N., Mouta, S., Brookhuis, K., deWaard, D. The interaction between driving and in-vehicle information systems: Comparison of results from laboratory, simulator and real-world studies. Transportation Research F, 8:135–146 (2005).
60. Schwalm, M. Die Schonung kognitiver Ressourcen durch die Nutzung auditiv-räumlicher Kommunikation in Navigationssystemen FR Psychologie. Saarbrücken: Universität des Saarlandes, (2006).
61. Shelton, A.L., Gabrieli, J.D.E. Neural correlates of individual differences in spatial learning strategies. Neuropsychology, 18:442–449 (2004).
62. Spence, C., Read, L. Speech shadowing while driving: On the difficulty of splitting attention between eye and ear. Psychological Science, 14:251–256 (2003).
63. Srinivasan, R., Jovanis, P.P. Effect of in-vehicle route guidance systems on driver workload and chioce of vehicle speed: Findings from a driving simulation experiment In Y.I. Noy (Ed.), Ergonomics and Safety of Intelligent Driver Interfaces (pp. 97–115). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc (1997).
64. Stahl, C., Haupert, J. Taking location modelling to new levels: A map modelling toolkit for intelligent environments. In Hazas, M., Strang, T., Krumm, J. (Eds.), Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3987 (pp. 74–85, vol. 3987/2006). Berlin Heidelberg, Munich, Germany: Springer-Verlag (2006).
65. Strayer, D.L., Drews, F.A., Johnston, W.A. Cell phone-induced failures of visual attention during simulated driving. Journal of Experimental Psychology: Applied, 9:23–32 (2003).
66. Streeter, L.A., Vitello, D., Wonsiewicz, S.A. How to tell people where to go: Comparing navigational aids. International Journal of Man-Machine Studies, 22:549–562 (1985).
67. Taylor, H.A., Tversky, B. Spatial mental models derived from survey and route descriptions. Journal of Memory and Language, 31:261–292 (1992).
68. Thorndyke, P.W., Hayes-Roth, B. Differences in spatial knowledge aquired from maps and navigation. Cognitive Psychology, 14:560–589 (1982).

69. Waller, D. Individual differences in spatial learning from computer-simulated environments. Journal of Experimental Psychology: Applied, 6:307–321 (2000).
70. Waller, D., Knapp, D., Hunt, E. Spatial representations of virtual mazes: The role of visual fidelity and individual differences. Human Factors, 43:147 (2001).
71. Wasinger, R., Krüger, A. Multi-modal interaction with mobile navigation systems. Special Journal Issue Conversational User Interfaces, it – Information Technology, 46(6):322–331 (2004).
72. Wasinger, R., Oliver, D., Heckmann, D., Braun, B., Brandherm, B., Stahl, C.: Adapting spoken and visual output for a pedestrian navigation system, based on given situational statements. In A. Hotho, G. Stumme (Eds.), LLWA 2003 (pp. 343–346), Karlsruhe, Germany: Lehren LernenWissen Adaptivität (2003).
73. Zacks, J.M., Michelon, P. Transformations of visuospatial images. Behavioral and Cognitive Neuroscience Reviews, 4:96–118 (2005).
74. Zimmer, H.D. The construction of mental maps based on a fragmentary view of physical maps. Journal of Eductional Psychology, 96:603–610 (2004).

Error-Induced Learning as a Resource-Adaptive Process in Young and Elderly Individuals

Nicola K. Ferdinand, Anja Weiten, Axel Mecklinger, and Jutta Kray

1 Introduction

Thorndike described in his law of effect [44] that actions followed by positive events are more likely to be repeated in the future, whereas actions that are followed by negative outcomes are less likely to be repeated. This implies that behavior is evaluated in the light of its potential consequences, and non-reward events (i.e., errors) must be detected for reinforcement learning to take place. In short, humans have to monitor their performance in order to detect and correct errors, and this allows them to successfully adapt their behavior to changing environmental demands and acquire new behavior, i.e., to learn.

For this type of learning, expectancies and their violation play a crucial role. While learning, we form expectancies about future events. When these expectancies are violated by something better or worse than expected, these prediction errors are detected by the error monitoring system, which in the following initiates adjustments of behavior to current task demands.

In this context, error monitoring is considered as a resource-adaptive cognitive process. The more efficient the error monitoring system, the better the learning of associative information. The primary aim of this chapter is to examine age-related changes in error monitoring, and by this, the ability of older adults to flexibly adapt to environmental changes. For this purpose, we will first introduce neural mechanisms underlying error monitoring. Then we will discuss whether error-induced learning is dependent on the intention to learn or not. Next, we will present empirical evidence that suggests age-related changes in error monitoring and learning. Finally, we will present a study which examines the above-mentioned processes.

N.K. Ferdinand (✉)
Experimental Neuropsychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany
e-mail: n.ferdinand@mx.uni-saarland.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_4,

2 Error Monitoring, ERN/Ne, and Dopamine

The basic mechanisms of reinforcement learning have been extensively studied in animals. A crucial neural system for this type of learning is the mesencephalic dopamine system. This system is composed of a small collection of nuclei, including the substantia nigra pars compacta and the ventral tegmental area, which are connected to brain structures involved in motivation and goal-directed behavior, e.g., the striatum, nucleus accumbens, and frontal cortex, including the anterior cingulate cortex (ACC; e.g., [3]). Schultz and colleagues recorded activity from mesencephalic dopamine cells in conditioning experiments with monkeys [41, 42] and found that the presentation of an unpredicted reward elicits a phasic response in dopamine neurons. Moreover, during learning the dopaminergic signal changes from the time a reward is delivered to when the conditioned stimulus is presented. Thus, the mesencephalic dopamine system can also become active in anticipation of a forthcoming reward. When an expected reward is not given, the mesencephalic dopamine neurons decrease their firing rate at the time the reward would normally have been delivered. Dopaminergic activity also falls below baseline when the monkey is presented with a stimulus that predicts punishment. Taken together, these findings from animal research indicate that the brain continuously makes predictions and then compares actual events and their consequences (reward or punishment) against those predictions. On the basis of these results, Schultz and colleagues proposed that dopamine neurons are sensitive to changes in the prediction of the "hedonistic value" of ongoing events: A positive dopamine signal is elicited when an event proves better than expected, and a negative one when an event proves worse.

In a model recently proposed by Holroyd and Coles [25], the findings from Schultz and colleagues [41, 42] have been adapted to humans and the mesencephalic dopamine system has been linked to error monitoring. Imagine you want to respond appropriately to a stimulus displayed on a screen by pressing a corresponding key. According to the Holroyd and Coles' model ([25]; outlined in Fig. 1), various motor controllers (e.g., the dorsolateral prefrontal cortex, the orbitofrontal cortex, or the amygdala) try to exert their influence over the motor system to solve this task in their own way. The anterior cingulate cortex (ACC), a brain region located in medial frontal cortex, acts as a control filter and regulates which of the controllers actually takes command of the motor system. To do this, the ACC must learn which controller is best suited for the task at hand. The model assumes that reinforcement learning signals conveyed via the mesencephalic dopamine system train the ACC to recognize the appropriate controller. During this process, the basal ganglia play an important role by learning to predict an event in terms of reward or punishment through experience: If an event is better than expected, i.e., an unexpected reward is given, phasic increases in mesencephalic dopamine activity are induced. If an event is worse than expected (e.g., you press the wrong response key although you know the correct one) the result is a phasic decrease in dopaminergic activity. In both cases, the ACC uses these predictive error signals to select and reinforce the motor controller that is most successful at performing the task at hand.

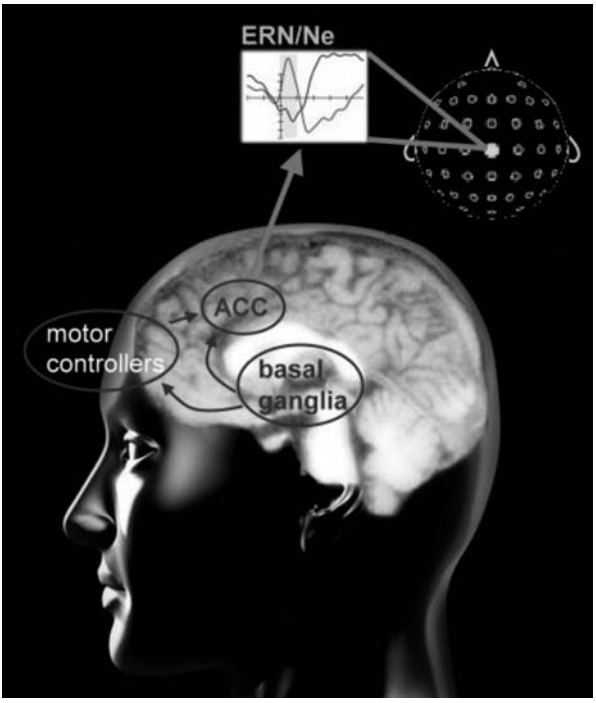


Fig. 1 According to the Holroyd and Coles model, various motor controllers try to exert their influence over the motor system to solve a task. The ACC acts as a control filter and regulates which of the them actually takes command. In order to recognize the appropriate controller the ACC is trained with reinforcement learning signals conveyed via the mesencephalic dopamine system

According to Holroyd and Coles [25], these phasic changes in dopaminergic activity can be monitored by scalp-recorded event-related potentials (ERPs). ERPs are averaged electroencephalogram (EEG) activity which is time-related to external and internal events. They therefore provide an excellent online-measure for the cognitive processes at work. When the dopaminergic reinforcement learning signal is lacking, an ERN/Ne (error-related negativity, ERN: [20]; or error negativity, Ne: [16]) can be measured at frontal central electrodes. This ERN/Ne is an ERP component elicited around the time an error is made. It can be observed in simple reaction time tasks (e.g., [16, 20]) as well as in recognition memory tasks [34] and coincides with response initiation, peaking roughly 80 ms afterwards. Its topographical maximum lies over fronto-central brain regions, and it is thought to be generated in the ACC (e.g., [46]). Furthermore, the ERN/Ne is sensitive to the degree of an error [4] and is influenced by its subjective significance [20, 21]. It is also elicited by error observation and by feedback that signals an error was committed [33], thus emphasizing the flexibility of the underlying error processing system.

To test their model, Holroyd and Coles [25] examined the ERN/Ne to erroneous responses and negative feedback as learning progressed throughout the course of a

probabilistic learning task, in which subjects had to learn stimulus-response mappings by trial and error, guided by feedback on each trial. They were able to show that when the feedback to correct or wrong responses was always valid, negative feedback elicited a feedback-ERN/Ne at the beginning of a learning block. As learning of the stimulus-response mappings progressed, the amplitude of the response-ERN/Ne became larger, while the amplitude of the feedback-ERN/Ne became smaller. They concluded that, as subjects learned the correct mappings, they relied less on the feedback and more on their own representation of what the response should be to determine the outcome of each trial (better or worse than expected).

Taken together, the mesencephalic dopamine system and the basal ganglia play a crucial role for reinforcement learning and are important parts of the human error monitoring system. This system is constantly monitoring for events contrary to our expectancies, e.g., committed errors, in order to optimize our behavior. We learn to solve a task by learning from those events that violate our expectancies, and the ERN/Ne is a highly valid indicator of the detection of these violations. The more resources are allocated to these processes, the more efficiently they can take place and the better the learning of associative information.

3 The Relevance of Learning Intention

As discussed above, the human brain learns by evaluating the results of our actions and this learning is driven by reward-related information carried to the ACC. However, this conclusion was derived from explicit learning experiments, in which subjects were instructed to learn and consequently had an intention to learn. This intention may have a key function in error-induced learning because it initiates an active search for expectancy violations (errors) and by this enhances the learning process.

In a previous ERP study, we investigated the role of error monitoring in implicit learning in a group of younger adults [18]. The focus of this study was on examining whether the detection of non-reward events and their implication for learning require an intention to learn or can occur without awareness. In other words, we were interested in whether it is possible to learn from errors that we are not aware of and therefore do not consciously assign processing resources to.

Implicit learning is defined as the acquisition of information without the intention to learn and without concurrent awareness that the material has been learned. In contrast, explicit learning is accompanied by both an intention to learn and awareness of the learned information ([5, 19, 38, 43]; for a review on implicit learning, see [9]). A paradigm frequently applied in implicit learning studies is the serial response time task [37]. In its original version, a stimulus is presented on a visual display in one of four possible locations. A specific button is assigned to each display location, and the participant's task is to quickly press the response button that corresponds to the location of the stimulus when it is displayed. When the sequence of stimulus locations follows a repeating pattern, reaction times decrease faster during the

experiment than if the sequence is random, indicating that this acceleration cannot only be based on practice effects but is also due to sequence learning. When the sequence is switched to a random sequence after prolonged practice with a repeating sequence, there is usually a marked increase in reaction times. Participants showing this pattern of results do not necessarily notice the presence of a repeating sequence nor are they able to verbalize their knowledge of the sequential structure. This suggests that they acquire this knowledge incidentally and without the assistance of conscious learning processes [12, 13, 37, 40, 39].

To examine the relation between the intention to learn and error monitoring we used a sequence learning paradigm with deviant stimuli inserted into an otherwise repeating sequence (for a similar procedure see [13, 39]). This allows to examine two different types of errors: First, subjects can press a wrong response button, which would be a "committed error." This type of error is most likely to be noticed by the subjects. Second, subjects can detect a stimulus that deviates from the regular sequence, which we refer to as a "perceived error" in the following. Since a sequence learning task can be administered under both explicit and implicit learning conditions, this strategy allows for investigating the influence of noticed and unnoticed perceived errors on learning under otherwise identical testing conditions.

In several learning studies, enhanced negative ERP components at about 200 ms (N200) have been reported for stimuli that violate participants' expectancies. For instance, using a contingency judgment task, Kopp and Wolff [29] showed that a stimulus that violated a predicted response elicited a fronto-centrally distributed N200 component. From their results they inferred that the N200 reflects brain events which register the mismatch between actual and expected sensory stimuli. Employing a sequence learning task, Rüsseler et al. [39] found an N200 component to deviant events, when participants learned intentionally. Even though not explicitly explored in these studies, it may be the case that these N200 components and the ERN/Ne reflect activity of a common neural generator (the ACC) initiated by input signaling that an event violates the participant's expectancy.

Following the above arguments, we predicted that deviant events would elicit a stimulus-related negative deflection about 200 ms after their presentation in the implicit and explicit learning condition (in the following referred to as N2b[1]). Moreover, we expected N2b amplitude to become more negative over the course of the experiment due to learning of the regular sequence. The idea was that deviant events would acquire the status of perceived errors during implicit and explicit learning. While performing the serial reaction time task, expectancies about upcoming events would be generated and evaluated on the dimension "better or worse than expected." Furthermore, the accuracy of this process should improve with learning and be reflected in a gradual increase in N2b amplitude as a function of learning. Our findings confirmed these predictions (see Fig. 2). Thus, perceived errors contribute to sequence learning even if subjects are not aware of them and do not consciously

[1] We will use the term N2b in the following to refer to the negativity after deviant events (perceived errors), to avoid confusion with the Mismatch Negativity (MMN) which is often denoted as N2a.

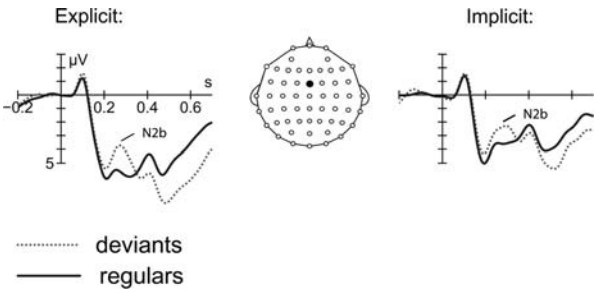


Fig. 2 N2b for explicit and implicit young subjects at electrode site FCz. Deviants are detected and perceived as errors, regardless of whether participants are consciously aware of them

assign cognitive control to process them. Although monitoring of perceived errors is more efficient when resources can be flexibly assigned to the task at hand (as indicated by a faster N2b development in explicit learning), it does still take place without conscious resource allocation in implicit learning [18].

In addition to the above-mentioned shared functional characteristics of committed and perceived errors (as shown by the ERN/Ne and N2b, respectively) we also found similar topographies for the two components. Thus, our data imply that the processing of committed and perceived errors may rely on the same neural mechanisms – albeit specialized for different sources of information (internal input for committed errors and visual input for perceived errors). The human brain learns by evaluating the results of our actions and this learning mechanism is driven by reward-related information carried to the ACC, initiating the cognitive control of motor and learning behavior. This assumption is supported by source-localization studies which show that the neural generators of the two components lie very close together in the medial frontal cortex, consistent with a common neural source in the ACC [24, 36].

4 Error-Induced Learning in the Elderly

The goal of this study was to investigate changes in explicit and implicit error monitoring in old age. As described above, the detection of non-reward events is assumed to drive explicit and implicit learning and is modulated by the mesencephalic dopamine system [25]. It is well known that aging is associated with pronounced changes in the mesencephalic dopamine system and in neural areas that receive input from this system, e.g., the prefrontal cortex [1, 2]. It has also been shown that the availability of dopamine D2 receptors in the striatum declines with age and that the availability of D2 receptors correlates with ACC glucose metabolism [47]. Thus, there is good evidence for age-related structural changes in the mesencephalic dopamine system. However, it remains unclear whether these changes have functional implications for error-induced learning in old age.

Following this line of thought, Nieuwenhuis and colleagues [35] have extended the neurocomputational model of Holroyd and Coles. Nieuwenhuis et al. [35]

conducted a probabilistic learning study very similar to that of Holroyd and Coles [25], and found a reduced response- and feedback-related ERN/Ne in older adults. They proposed that the reduced ERN/Ne found in older adults is caused by a weakened reinforcement signal from the mesencephalic dopamine system to the ACC and that this results in impaired learning from events that are worse than expected. In support of this view, there is evidence from other explicit learning studies showing that an impaired mesencephalic dopamine system in old age affects error processing and that this impairment is also visible in reduced error-related ERP components. For instance, Falkenstein et al. [17] found reduced ERN/Ne amplitudes in response to committed errors in elderly participants in a choice-reaction task. Interestingly, this effect was not due to a general reduction of ERP amplitudes in the elderly since other ERP components were not affected. In addition, Eppinger et al. [15] were able to show that learning is also reflected in a response-locked positivity for correct trials and a feedback-locked positivity after positive feedback. The mechanism reflected in the latter ERP components appears to be especially important for learning in older adults, who seem to rely more on positive than negative feedback during learning.

So, on the one hand, an age-impaired mesencephalic dopamine system should affect error-induced learning, and we should be able to observe this in impaired learning performance as well as in attenuated error-related ERP components. On the other hand, there is evidence for intact implicit learning in the elderly. For example, Howard and Howard [26, 27] assigned younger and older adults to a sequence learning task. They were able to show that, although overall older adults produced longer reaction times, both young and elderly adults were equally disrupted by the switch from a repeating sequence to random series of items. To get an estimate of explicit sequence learning, subjects were asked to predict the next item of a learned sequence, and this revealed reliable age-related deficits. In contrast, the implicit learning measure did not show an age-related deficit. In addition, there is some evidence to suggest that the neural substrates supporting implicit learning may be largely unaffected by aging and by the kinds of dysfunctions that compromise explicit learning, such as Korsakoff's syndrome (e.g., [11]). However, age differences have been reported for implicit learning tasks when the task is relatively difficult or when the older subjects are of lower intellectual ability (e.g., [8, 22]). It therefore remains an open issue on how implicit and explicit error monitoring and error-induced learning are affected by old age.

5 Methods and Procedure

The main issues addressed in this study were (a) modulations in error monitoring with old age and (b) their influence on explicit and implicit learning processes. Therefore, we examined 45 elderly participants (mean age = 68.6 years, age range = 64–75 years). 37 young subjects (mean age = 21.1 years, age range = 18–27 years) from our previous study [18] served as a control (see Table 1).

We presented letters one-by-one on a computer screen. Subjects responded to the letters by pressing a corresponding response key as quickly as possible.

Table 1 Description of sample

	Young		Elderly	
	Explicit	Implicit	Explicit	Implicit
N	19	18	19	20
Gender	9 female/10 male	8 female/10 male	9 female/10 male	9 female/11 male
Mean age	21.32	20.61	68.11	68.85
Age range	19–24	18–27	65–74	64–75
Operation span	16.05	14.56	10.00	13.33
Digit symbol	64.58	64.89	47.89	45.13

The letters were presented in either a regular, irregular, or random sequence. In regular sequences, letters were presented according to a fixed eight-letter sequence (CBADBCDA). In irregular sequences, one letter in the regular sequence was replaced by a letter that otherwise had not occurred at that position within the sequence.

Each subject performed two blocks of regular and irregular sequences, drawn in a random order. A random-sequence block followed each regular/irregular block. This design allowed for two operational definitions of sequence learning: First, changes in reaction times to regular and irregular sequences over the course of the entire experiment, and second, the difference in reaction times between regular stimuli at the end of a block and the following random stimuli (a measure that is relatively free from practice effects).

Young and elderly subjects were randomly divided into two groups, respectively, and assigned to either an explicit or implicit sequence learning task. Participants in the two explicit learning groups were told that the letters would appear in a mostly repeating sequence and that they should learn this sequence in order to improve their performance. In contrast, the presence of a sequence was not revealed to the implicit learning groups (for a more detailed description of the procedure see [18]).

A variety of memory tests were applied to ensure that subjects in the implicit learning groups had not become aware of the repeating sequence during the experiment (for a detailed description of these tests see [18]). We excluded three elderly participants from our implicit-learning group because their performance on our memory tasks indicated they had acquired explicit sequence knowledge. Additionally, one elderly subject from the implicit group had to be excluded from the analyses because of technical artifacts during EEG recording and two more (one from the implicit and one from the explicit group) because they showed more than 50% missing values. Consequently, all statistical analyses were based on 39 older (20 implicit and 19 explicit) and 37 younger (18 implicit and 19 explicit) participants.

Also, in order to examine the relationship between error monitoring and the level of cognitive fitness, all participants completed an operation span test [45, 14] and a digit symbol test (adapted from Wechsler [48]) in a separate session. The operation span is a measure of working memory capacity, and the digit symbol test is thought to reflect perceptual speed of processing. Both, working memory capacity and processing speed reflect abilities of the domain of fluid intelligence and are known to decline in old age (for recent reviews see [10, 31, 32]).

The experimental procedure described above offers the opportunity to examine two types of error monitoring: one involved in response-related processing and the other involved in stimulus-related processing. Response-related processing deals with the detection of wrong responses (i.e., *committed errors*) and is likely to be noticed by all subjects irrespective of learning condition. The second type of error monitoring focuses on the detection of deviant stimuli. In the context of the regular sequence, subjects develop expectancies about the next stimulus. This formation and evaluation of expectancies is an important and for the explicit group (who is requested to detect and learn the repeating sequence) even necessary part of sequence learning. Since a deviant stimulus in our task can never be predicted, it will always be perceived as an unfavorable event, an error, in the context of sequence learning. Thus, we consider the detection of these stimuli a *perceived error*. This second error monitoring type is especially important because perceived errors (as opposed to committed errors) can be processed both with and without awareness. Since these perceived errors occur for both learning conditions, they allow us to directly compare implicit and explicit error monitoring processes and their resulting effects on learning.

In accordance with the above studies, we expected sequence learning to be evidenced by decreasing reaction times to regular stimuli over the course of the experiment, whereas reaction times for deviant stimuli should not show this pattern. Also, reaction times to regular stimuli should be shorter than those to random stimuli after subjects have learned the repeating sequence. Both effects should be more pronounced for the explicit learning condition than for the implicit learning condition. Furthermore, we expected impaired sequence learning for explicit elderly learners, whereas implicit learning should be less affected by age.

As for the response-related ERPs, we predicted an ERN/Ne to incorrect responses, signaling the detection of committed errors. Because of a weakened dopamine system, older participants should show reduced ERN/Ne amplitudes following committed errors as compared to younger participants. With regard to the processing of deviants, we predicted that they acquire the status of perceived errors and elicit an N2b component, regardless of whether subjects are aware of them. While performing the task, expectancies about upcoming events are generated and evaluated on the dimension “better or worse than expected” and this is reflected in the presence of the N2b. Since implicit and explicit error monitoring processes are considered to rely on the same neural mechanisms [18, 24, 36] we also expected smaller N2b amplitudes for older than for younger subjects.

6 Results

6.1 Reaction Times

Reaction times to regular and random stimuli were compared in order to examine whether participants successfully learned the structure of the sequence (Fig. 3).

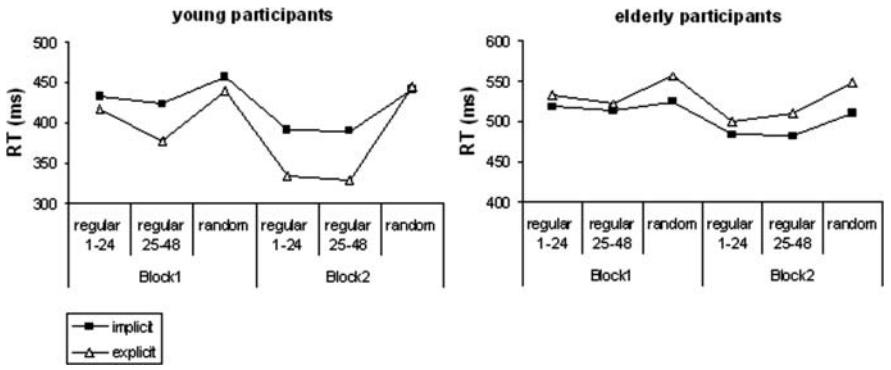


Fig. 3 Mean reaction times for young and elderly participants for correct responses to regular and random stimuli in both learning conditions (irregular sequences not included). Learning was measured as the difference in reaction times between regular and random stimuli. In the elderly, the same amount of learning was obtained in the explicit and implicit condition, while for young participants learning-related differences in response times were more pronounced in the explicit than in the implicit condition

A repeated measures analysis of variance (ANOVA) with the between-subjects factors Age (young/elderly) and Learning Condition (explicit/implicit) and the within-subjects factors Stimulus Type (regular stimuli from the second half of a block/random stimuli) and Block (block1/block2) showed that older participants had generally slower reaction times than younger participants (Age: $F(1, 72) = 90.04$, $p < 0.01$). An interaction between Age and Learning Condition ($F(1, 72) = 6.35$, $p = 0.01$) indicated that for young subjects, reaction times in the explicit condition were faster than in the implicit condition ($F(1, 36) = 4.11$, $p = 0.05$), while this was not the case for older subjects ($p = 0.12$). Both age groups learned the sequential structure of the material as reflected in faster reaction times to regular than to random stimuli (Stimulus Type: $F(1, 72) = 205.81$, $p < 0.01$). This reaction time gain was less pronounced in the older than in the younger age group (Age×Stimulus Type: $F(1, 72) = 32.58$, $p < 0.01$). Reaction times for both age groups were faster in the second block (Block: $F(1, 72) = 23.92$, $p < 0.01$), and the reaction time gain was also larger in the second half of the experiment (Stimulus Type×Block: $F(1, 72) = 24.54$, $p < 0.01$), especially for the younger subjects (Age×Stimulus Type×Block: $F(1, 72) = 8.01$, $p < 0.01$). Additionally, explicit learners benefited more from sequence learning than implicit learners as revealed by an interaction between Learning Condition and Stimulus Type ($F(1, 72) = 23.94$, $p < 0.01$). However, this interaction was observed only in young adults, and there were no reaction time differences between explicit and implicit learners for the senior adults (Age×Learning Condition×Stimulus Type: $F(1, 72) = 5.38$, $p = 0.02$). Finally, the four-way interaction between Age, Learning Condition, Stimulus Type, and Block was also significant ($F(1, 72) = 6.09$, $p = 0.02$).

In order to compare the changes in reaction times for correct answers in response to regular stimuli with those in response to deviant stimuli, we divided the

experiment into four bins, each containing 24 regular and 24 irregular sequences. An ANOVA with the between-subjects factors Age (young/elderly) and Learning Condition (explicit/implicit) and the within-subjects factors Stimulus Type (regular/deviant) and Bin (1/2/3/4) again showed slower reaction times for the elderly (main effect Age: $F(1, 72) = 95.25,\ p < 0.01$), particularly in the explicit learning condition (Age×Learning Condition: $F(1, 72) = 5.47,\ p = 0.02$). Reaction times were generally faster to regular than to deviant stimuli (Stimulus Type: $F(1, 72) = 231.73,\ p < 0.01$), especially for young participants (Age×Stimulus Type: $F(1, 72) = 18.47,\ p < 0.01$) and for the explicit group (Learning Condition × Stimulus Type: $F(1, 72) = 17.31,\ p < 0.01$). During the course of the experiment reaction times became faster (Bin: $F(3, 216) = 25.96,\ p < 0.01,\ \varepsilon = 0.64$), mainly in response to regular stimuli as shown by an interaction between Stimulus Type and Bin ($F(3, 216) = 13.45,\ p < 0.01,\ \varepsilon = 0.92$). In addition, we found an interaction between Stimulus Type, Bin, and Learning Condition ($F(3, 216) = 7.75,\ p < 0.01,\ \varepsilon = 0.92$), reflecting greater reaction time decreases in the explicit condition, especially to regular stimuli, and an interaction between Stimulus Type, Bin, and Age ($F(3, 216) = 7.89,\ p < 0.01,\ \varepsilon = 0.92$) showing that reaction times in the group of young subjects became faster only to regular stimuli.[2]

6.2 Error Rates

To examine the extent to which learning was visible in the error rates, we conducted the same analyses for the accuracy data. An ANOVA with the factors Age (young/elderly), Learning Condition (explicit/implicit), Stimulus Type (regular/random), and Block (block1/block2) showed that there was no general difference in error rate between younger and older participants ($p = 0.24$), but that older participants made more errors in the implicit learning condition (Age × Learning Condition: $F(1, 72) = 4.62,\ p = 0.03$), and this effect was even more pronounced in the second half of the experiment (Age×Learning Condition×Block: $F(1, 72) = 4.49,\ p = 0.04$). Moreover, subjects committed more errors in response to random than to regular stimuli, indicating that sequence learning had taken place (Stimulus Type: $F(1, 72) = 76.16,\ p < 0.01$). Overall, this learning effect was marginally smaller for elderly adults (Age×Stimulus Type: $F(1, 72) = 3.38,\ p = 0.07$), which was due mainly to differences between the age groups in the explicit learning condition (Age× Learning Condition × Stimulus Type: $F(1, 72) = 3.82,\ p = 0.05$), reflecting impaired sequence learning for the elderly in that condition (Fig. 4).

An ANOVA with the factors Age (young/elderly), Learning Condition (explicit/implicit), Stimulus Type (regular/deviant), and Bin (1/2/3/4) served to examine changes in error rates over the course of the experiment. All in all, elderly

[2] In order to ensure that the described interactions containing the factor Age were not only due to reaction time differences between the age groups, reaction times were log-transformed [28, 30] and all statistical analyses were repeated. However, the results remained unchanged by this procedure.

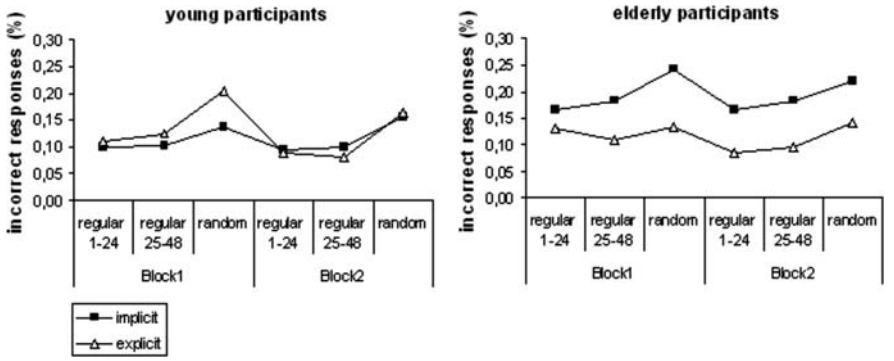


Fig. 4 Error rates for young and elderly participants to regular and random stimuli and for both learning conditions (irregular sequences not included). All groups make more errors in response to random than to regular stimuli

subjects made fewer errors than younger subjects in the explicit condition, while there was no such difference in the implicit condition, as revealed by an interaction between Age and Learning Condition ($F(1, 72) = 8.52$, $p < 0.01$). Additionally, more errors were made in response to deviants than to regulars (Stimulus Type: $F(1, 72) = 206.65$, $p < 0.01$), primarily in the explicit condition (Learning Condition $\times$ Stimulus Type: $F(1, 72) = 14.96$, $p < 0.01$). This effect was also more pronounced for younger adults (Age $\times$ Stimulus Type: $F(1, 72) = 21.94$, $p < 0.01$) and again mainly in the explicit condition (Age $\times$ Learning Condition $\times$ Stimulus Type: $F(1, 72) = 7.98$, $p < 0.01$). Furthermore, the error rates increased in a linear manner over the course of the experiment (Bin: $F(3, 216) = 8.59$, $p < 0.01$, $\varepsilon = 0.83$), and this increase was more pronounced for younger adults than for older adults (Age $\times$ Bin: $F(3, 216) = 4.67$, $p < 0.01$, $\varepsilon = 0.83$), particularly in the explicit condition (Age $\times$ Learning Condition$\times$Bin: $F(3, 216) = 2.95$, $p = 0.04$, $\varepsilon = 0.83$). Moreover, this increase in error rates was due to deviant stimuli (Stimulus Type $\times$ Bin: $F(3, 216) = 22.46$, $p < 0.01$). Finally, the results showed significant interactions between Stimulus Type, Bin, and Age ($F(3, 216) = 8.0$, $p < 0.01$, $\varepsilon = 0.99$), Stimulus Type, Bin, and Learning Condition ($F(3, 216) = 3.0$, $p = 0.03$, $\varepsilon = 0.99$), and Stimulus Type, Bin, Learning Condition, and Age ($F(3, 216) = 4.12$, $p < 0.01$, $\varepsilon = 0.99$). Contrasts revealed that the two age groups demonstrated different error patterns: For younger subjects, learning was reflected in error rates to deviants that increased over the course of the experiment, and this effect was larger for explicit young learners. However, elderly explicit learners showed a decrease in their error rates in response to regular stimuli.

6.3 Speed-Accuracy Trade-Off

Elderly subjects often apply different speed-accuracy trade-off strategies than younger subjects. They trade speed for higher accuracy (e.g., [7, 6]). In this study, we found that older adults made less errors in the explicit than in the implicit

learning condition. Also, the reaction times in the explicit elderly group were slower than those in the implicit elderly group. This finding was unexpected, and we wondered whether it could be due to different speed-accuracy trade-off strategies. More specifically, we wondered whether the focus of the explicit elderly group was on responding as accurately as possible, while the other groups focused on responding as quickly as possible.

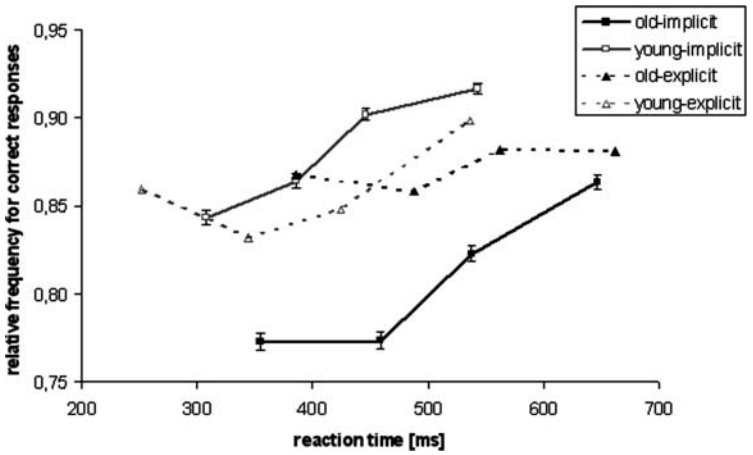


Fig. 5 Speed-accuracy trade-offs for young and elderly explicit and implicit participants. All the groups except for the explicit elderly showed a speed-accuracy trade-off; the faster the reaction times, the more errors were made. In contrast to that, the explicit elderly showed the same error rate in every reaction time quartile

To control for potential age and learning group differences in speed-accuracy trade-offs, we divided reaction times for each subject into quartiles, calculated the mean accuracy for each, and then averaged them across subjects for each age and learning group (Fig. 5). An ANOVA with the factors Age (young/elderly), Learning Condition (explicit/implicit), and Quartile (1/2/3/4) revealed that the faster the responses, the more errors were made (Quartile: $F(3, 216) = 21.56$, $p < 0.01$, $\varepsilon = 0.60$). Also, we found an interaction between Age and Learning Condition ($F(1, 72) = 4.10$, $p = 0.05$) that indicated that the explicit and implicit learning condition differed marginally for older adults ($F(1, 37) = 3.06$, $p = 0.09$) and that both younger and older adults differed in the implicit learning condition ($F(1, 36) = 3.81$, $p = 0.05$). Finally, the analysis showed an interaction between Quartile and Learning Condition ($F(3, 216) = 5.10$, $p = 0.01$, $\varepsilon = 0.60$). Post-hoc contrasts revealed that in the implicit learning condition, error rate increased with faster responding (error $\text{rate}_{Q1} = \text{error rate}_{Q2} > \text{error rate}_{Q3} > \text{error rate}_{Q4}$; lower quartile numbers indicate faster reaction times). For the explicit learning group we also found an increased error rate with faster responding but only from the second to the fourth quartile (error $\text{rate}_{Q1} < \text{error rate}_{Q2}$, error $\text{rate}_{Q1} > \text{error rate}_{Q4}$, error $\text{rate}_{Q2} > \text{error rate}_{Q3} > \text{error rate}_{Q4}$). Additionally, the three-way interaction between Age, Learning Condition, and Quartile was marginally significant ($F(3, 216) = 2.63$, $p = 0.08$, $\varepsilon = 0.60$),

indicating that all the groups except for the explicit elderly showed a speed-accuracy trade-off; the faster the reaction times, the more errors were made. In contrast to that, the explicit elderly showed the same low error rate in every reaction time quartile. This finding will be evaluated further below.

6.4 ERPs for Committed Errors

As can be seen in Fig. 6, erroneous responses elicited a pronounced ERN/Ne with a maximum amplitude at Cz. An ANOVA with the factors Age (young/elderly), Group (explicit/implicit), and Response (incorrect/correct) in the time window from 0 to 100 ms revealed that the mean amplitude for incorrect responses was more negative than that for correct responses (Response: $F(1, 72) = 90.51$, $p < 0.01$) and that mean amplitudes for incorrect responses were more negative for the younger than for the older adults (Age×Response: $F(1, 35) = 4.83$, $p = 0.03$).

The topography of the ERN/Ne was similar for younger and older explicit and implicit learners. An ANOVA with factors Age (young/elderly), Learning Condition (explicit/implicit), Anterior–posterior (frontal/central/parietal/occipital

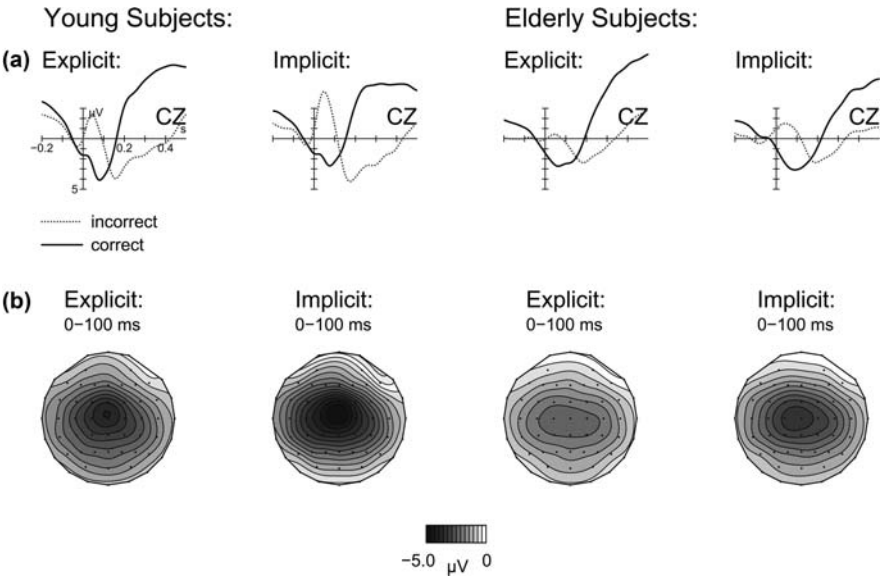


Fig. 6 (**a**) Response-locked ERP waveforms elicited by correct and erroneous answers at electrode site Cz displayed separately for young and elderly implicit and explicit learners. All groups show a pronounced ERN/Ne to erroneous answers, the component being smaller for elderly subjects. Even though the figure suggests that the ERN/Ne is smaller in the explicit as compared to the implicit conditions, this was not significant in the overall analysis. (**b**) Topographic difference maps of the ERN/Ne for young and elderly implicit and explicit learners (incorrect – correct answers) in the time window 0–100 ms

electrode sites), and Lateralization (left/middle/right)[3] confirmed that there was no main effect or significant interaction involving the factor Learning Condition. However, we found an interaction between Anterior–posterior, Lateralization, and Age ($F(6, 432) = 2.51$, $p = 0.05$, $\varepsilon = 0.57$), indicating a broader ERN/Ne scalp distribution for the elderly.

6.5 ERPs for Perceived Errors

Mean amplitudes of stimulus-locked ERP waveforms were examined in the time window of the N2b, 220–320 ms after stimulus onset at electrode FCz. The ANOVA included the factors Age (young/elderly), Learning Condition (explicit/implicit), and Stimulus Type (regular/deviant). The mean amplitude was larger for the explicit than for the implicit learning condition (Learning Condition: $F(1, 72) = 4.83$, $p = 0.03$) and it was significantly more negative for deviants than for regular stimuli (Stimulus Type: $F(1, 72) = 13.92$, $p < 0.01$; see Fig. 7). However, this effect interacted with age (Age × Stimulus Type: $F(1, 72) = 20.80$, $p < 0.01$). Post-hoc analyses revealed that the larger N2b to deviants than to regulars was only found for the group of younger adults ($F(1, 36) = 28.51$, $p < 0.01$) but not for the older adults ($p = 0.53$).

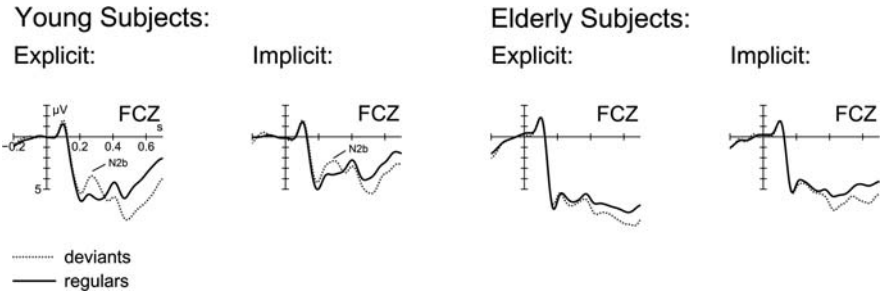


Fig. 7 Stimulus-locked N2b waveforms elicited by regular and deviant stimuli at electrode site FCz displayed separately for young and elderly implicit and explicit learners. While young subjects from both learning conditions display a pronounced N2b, there is no such effect for the groups of elderly subjects

The finding that there was no N2b for the elderly in the overall analysis might be due to poor sequence learning in those participants. To check this possibility, we grouped our subjects according to their learning performance[4] and compared

[3] For topographical analysis of the ERN/Ne the following electrodes were used: F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, O1, Oz, and O2 (according to the international 10–20 system).

[4] Learning performance was measured as reaction time gain in response to regular vs. random stimuli. Comparable sequence learning performance between good elderly (explicit: $N = 9$, mean reaction time gain = 56 ms; implicit: $N = 10$, mean reaction time gain = 40 ms) and poor young learners (explicit: $N = 9$, mean reaction time gain = 61 ms; implicit: $N = 9$, mean reaction time gain = 33 ms) was found for both the explicit and the implicit learning condition.

good elderly learners with poor young learners. Then we analyzed the relationship between learning performance and N2b amplitude with the same ANOVA as before. We found a main effect for Stimulus Type, indicating that mean amplitude was more negative for deviants ($F(1, 33) = 7.34,\ p = 0.01$) and an interaction between Age and Stimulus Type ($F(1, 33) = 13.29,\ p < 0.01$). Post-hoc contrasts confirmed an effect of Stimulus Type for the poor sequence learners of the younger age group ($F(1, 17) = 27.36,\ p < 0.0001$), but not for the good sequence learners of the older age group ($p = 0.55$). These results show that the absence of the N2b in the older group was not due to poor learning performance.

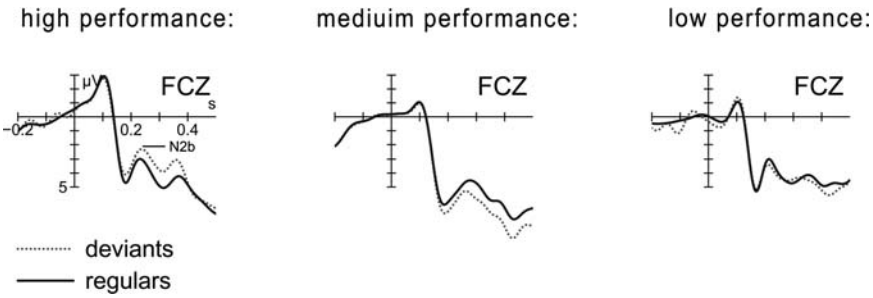


Fig. 8 Stimulus-locked N2b waveforms elicited by regular and deviant stimuli at electrode sites FCz for elderly implicit learners according to performance in the operation span task. Elderly participants with high performance in the operation span task show an N2b to deviant stimuli

As the absence of the N2b effect in older adults was quite surprising, we additionally analyzed whether this effect was modulated by differences in the level of cognitive functioning between younger and older adults. Thus, we assumed that only older adults with a higher level of cognitive functioning as measured by working memory capacity (operation span; see Table 1) have a monitoring system efficient enough to detect perceived errors. To examine this, we divided the groups of older adults by their performance in the operation span task. We then conducted an ANOVA for the older adults with the factors Performance (low/medium/high operation span value), Learning Condition (explicit/implicit), and Stimulus Type (deviant/regular). Importantly, the results showed a significant three-way interaction ($F(2, 33) = 3.9,\ p = 0.03$). Implicit learners, who had performed well in the operation span, also showed an N2b to deviant stimuli (see Fig. 8). This N2b was not obtained for elderly with low or medium operation span. A different pattern was found for explicit learners. In this group, mean amplitudes were more negative in response to regular stimuli than to deviants independently of working-memory span. When we divided older adults by their performance in the digit symbol task, no such effect was found.

7 Discussion

The main goal of this study was to investigate error monitoring processes during implicit and explicit sequence learning and how those processes are modulated by

old age. In particular, we were interested in whether errors that subjects are unaware of influence learning and whether this changes in old age. The idea was that our error monitoring system continuously checks for events that violate our expectancies, that is, it evaluates events on the dimension "better/worse than expected." Recent models and empirical evidence [25, 41, 42] suggest that if an event proves better than expected and reward is anticipated, phasic increases in mesencephalic dopamine activity occur. If an event is worse than expected, the result is a phasic decrease in dopaminergic activity. By this process, committed and perceived errors are detected (as indicated by the ERN/Ne and N2b, respectively), irrespective of whether they are consciously noticed, and are used to optimize task performance. Since error detection is a resource-adaptive cognitive process and dependent on an intact dopamine system, but implicit learning processes appear to be spared in older adults, it is still an open issue how error-induced explicit and implicit learning is modulated in old age. In order to examine this, we conducted a serial reaction time task under an explicit and an implicit learning condition with young and elderly adults.

Although the elderly generally displayed slower reaction times, both age groups responded more slowly to random than to regular sequences. We also found that correct responses to stimuli from regular sequences became faster over the course of the experiment. These effects were present in all groups, but were more pronounced in the group of explicit young subjects. Thus, we inferred that sequence learning took place in both learning conditions and in both age groups, but was impaired in older participants from the explicit group. This was concluded because the explicitly learning older participants did *not* show larger learning effects than older participants from the implicit group, and thus did not benefit from explicit learning instructions as the younger subjects did.

Evidence of regular-sequence learning also appeared in the accuracy data. Overall, subjects made fewer errors in response to stimuli from the regular sequence than to those from random sequences. This effect was smallest for older adults from the explicit group, again reflecting impaired sequence learning in that group. With increasing time on task, we found different learning patterns for younger and older participants. For younger adults, errors in response to deviant stimuli increased over the course of the experiment, while errors to regular stimuli did not, and this effect was even larger for explicit young learners. In contrast, older subjects from the explicit group showed a decrease in their error rate in response to regular stimuli, while those from the implicit group showed no change in learning-related error rates. Also, older adults from the explicit group made less errors than those from the implicit group. This was unexpected and we wondered whether it might be due to a different speed-accuracy trade-off in the explicit elderly. We found that all the learning groups except for the explicit elderly showed the expected speed-accuracy trade-off; the faster the reaction times, the more errors were made. In contrast to that, the explicit elderly showed the same error rate in every reaction time quartile. Although the respective effects were only marginally significant they may account for this group's learning impairments. A possible reason for this finding could be that the explicit learning condition was too demanding for older participants. The

explicit groups have two tasks: detect and learn the regular sequence and respond as quickly and accurately as possible. The timing constraints of this study, with a response interval of 800 ms, may have made this too difficult for the elderly. While the other groups kept to the instructions and tried to respond as quickly and as accurately as possible, the explicit elderly tried to switch to a response strategy that emphasized accurate responses over fast responses. With the timing constraints in our design the amount of reaction time that can be traded for higher accuracy is limited (there were only 800 ms to respond), and so slower responding beyond that point did not result in improved accuracy for the explicit elderly. The strategy of the explicit elderly also means that they had to process stimuli more on a trial-by-trial basis to ensure high accuracy, and this may explain the impaired sequence learning in that group. Interestingly, in support of these considerations, elderly subjects from the explicit group also showed the slowest reaction times.

Erroneous responses elicited an ERN/Ne with maximum amplitude at central electrodes for young and elderly learners in both learning conditions, signaling the detection of a committed error. In line with our predictions, this effect was smaller for the elderly. Since ERN/Ne topography was more focused for young learners but otherwise did not differ between learning conditions, the component seems to be generated by the same underlying brain mechanisms in both cases. The smaller ERN/Ne for older as compared to younger adults supports the idea that error monitoring is a dopamine-related process and that impairments in the mesencephalic dopamine system (such as those present in old age) result in poorer error processing.

Consistent with our prediction that deviants acquire the status of perceived errors, we found an N2b to deviant stimuli. However, this effect was found only in young participants, and overall, no N2b was found in the elderly. For young subjects, this component was present for the implicit and the explicit learning group, implying that deviants are detected regardless of whether subjects are aware of them. Moreover, for the implicit group the N2b gradually increased during the course of the experiment, indicating a direct relationship between the acceleration of response times to regular stimuli and the development of the N2b to deviants: The faster subjects responded, the more negative was the N2b amplitude to deviant stimuli (cf. [18]). This result shows that the N2b is related to a gradual development of knowledge about the sequence structure and accompanies the formation of expectancies about the next stimulus. The expected and actual stimulus are then compared, and the event is evaluated as "better or worse than expected" to avoid future prediction errors, thus enabling learning.

We examined several rationales for the absence of the N2b in the older group. First, we compared young and senior adults with matched learning performance to ensure that the absence of an N2b in the older group was not due to weaker sequence learning performance. Yet, we still found no negativity in response to deviant stimuli for the elderly. Second, we searched for a relationship between N2b amplitude and working memory capacity in the elderly, as measured by the operation span task. Working memory capacity is an important characteristic of high cognitive functioning in the domain of fluid intelligence that is known to decline with age (e.g., [23];

for reviews, see [31, 32]) and may contribute to sequence learning. We found that implicit elderly learners who performed well in the operation span task showed an N2b to deviants. We did not find such a relationship for elderly adults from the explicit group. Since explicit elderly subjects also showed poorer operation span performance in general, it cannot be unambiguously decided whether the absence of an N2b effect was due to this group's low working memory capacity, or to being overstrained by the task demands (as indexed by the different speed-accuracy trade-off function), or a combination of both factors.

To conclude, for young subjects error monitoring processes take place during implicit and explicit learning and also play a crucial role for learning itself: While performing a task, expectancies about upcoming events are generated and evaluated on the dimension "better or worse than expected." The accuracy of this process improves with learning and this is reflected in gradual changes in error-related ERP components such as the response and feedback ERN/Ne [25] and the N2b [18] as a function of learning. In this way, even errors subjects are not aware of can contribute to sequence learning although they do not consciously assign cognitive control to detecting and processing them. Although error monitoring is more efficient when resources can flexibly be assigned to the task at hand (as in explicit learning), it does still take place without intentional resource allocation (as in implicit learning).

In accordance with the assumption that error monitoring can be perceived of as a cognitive resource that declines with age and that relies on a dopamine-related process, we found reduced ERN/Ne amplitudes to committed errors in older participants. In response to perceived errors (deviant events), we found no overall N2b in our older groups. However, for implicitly learning older adults we found a reliable relationship between N2b amplitude and working memory capacity. This relationship was not found for elderly participants from the explicit learning condition, the group that also showed the largest learning impairments. It is possible that this was due to lower working memory capacity and too high task demands in that group. Still, both of these accounts speak in favor of a weakened dopamine system in old age that is probably related to impaired error monitoring and consequently weaker learning performance. However, if sequence learning was solely relying on dopaminergic reinforcement signals, it has to be explained why elderly subjects from the implicit group showed behavioral learning effects almost comparable to those of the young implicit group. There are two possible explanations for this. The first one is that implicit learning is at least partly based on different learning mechanisms than explicit learning and depends on brain structures which are spared in old age (for an overview see [11]). The second is that the dopamine-mediated learning system is degraded in older participants with low working memory capacity or with different processing strategies. Further studies will be required to decide between these two interpretations.

Acknowledgments This research was supported by the Deutsche Forschungsgemeinschaft (grant SFB 378, EM 2). We wish to thank Anita Althausen and Martina Hubo for their support during data collection and Ben Eppinger for helpful comments.

References

1. Bäckman, L., Ginovart, N., Dixon, R.A., Robins Wahlin, T.-B., Wahlin, Å., Halldin, C., Farde, L. Age-related cognitive deficits mediated by changes in the striatal dopamine system. American Journal of Psychiatry, 157:635–637 (2000).
2. Bäckman, L., Nyberg, L., Lindenberger, U., Li, S.C., Farde, L. The correlative triad among aging, dopamine, and cognition: Current status and future prospects. Neuroscience and Biobehavioral Reviews, 30:791–807 (2006).
3. Berger, B., Gaspar, P., Verney, C. Dopaminergic innervation of the cerebral cortex: Unexpected differences between rodents and primates. Trends in Neurosciences, 14:21–27 (1991).
4. Bernstein, P.S., Scheffers, M.K., Coles, M.G.H. "Where did I go wrong?" A psychophysiological analysis of error detection. Journal of Experimental Psychology: Human Perception and Performance, 21:1312–1322 (1995).
5. Berry, D.C. Implicit learning: Twenty-five years on. A tutorial. In C. Umilta, M. Moscovitch (Eds.), Attention and Performance XV: Conscious and Nonconscious Information Processing (pp. 755–782). Cambridge, MA: MIT Press (1994).
6. Botwinick, J. Aging and Behavior. New York: Springer (1984).
7. Cerella, J. Aging and information-processing rate. In J.E. Birren, K.W. Schaie (Eds.), Handbook of the Psychology of Aging (pp. 201–221). San Diego: Academic Press (1990).
8. Cherry, K.E., Stadler, M.A. Implicit learning of a nonverbal sequence in younger and older adults. Psychology and Aging, 10:379–394 (1995).
9. Cleeremans, A., Destrebecqz, A., Boyer, M. Implicit learning: News from the front. Trends in Cognitive Sciences, 2:406–416 (1998).
10. Craik, F.I.M., Bialystok, E. Cognition through the lifespan: Mechanisms of change. Trends in Cognitive Sciences, 10:131–138 (2006).
11. Curran, T. On the neural mechanisms of sequence learning. Psyche [On-line serial], 2(12), Available at URL: http://psyche.csse.monash.edu.au/v2/psyche-2–12-curran.html, (2006).
12. Destrebecqz, A., Cleeremans, A. Can sequence learning be implicit? New evidence with the process dissociation procedure. Psychonomic Bulletin and Review, 8:343–350 (2001).
13. Eimer, M., Goschke, T., Schlaghecken, F., Stürmer, B. Explicit and implicit learning of event sequences: Evidence from event-related brain potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22:970–987 (1996).
14. Engle, R.W. Working memory capacity as executive attention. Current Directions in Psychological Science, 11:19–23 (2002).
15. Eppinger, B., Kray, J., Mock, B., Mecklinger, A. Better or worse than expected? Aging, learning, and the ERN. Neuropsychologia, 46:521–539 (2008).
16. Falkenstein, M., Hohnsbein, J., Hoorman, J., Blanke, L. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C.H.M. Brunia, A.W.K. Gaillard, A. Kok (Eds.), Psychophysiological Brain Research (pp. 192–195). Tilburg: Tilburg University Press (1990).
17. Falkenstein, M., Hoorman, J., Hohnsbein, J. Changes of error-related ERPs with age. Experimental Brain Research, 138:258–262 (2001).
18. Ferdinand, N.K., Mecklinger, A., Kray, J. Error and deviance processing in implicit and explicit sequence learning. Journal of Cognitive Neuroscience, 20:629–642 (2008).
19. Frensch, P.A. One concept, multiple meanings. On how to define the concept of implicit learning. In M.A. Stadler, P.A. Frensch, (Eds.), Handbook of Implicit Learning (pp. 47–104). Thousand Oaks: Sage (1998).
20. Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., Donchin, E. A neural system for error detection and compensation. Psychological Science, 4:385–390 (1993).
21. Hajcak, G., Moser, J.S., Yeung, N., Simons, R.F. On the ERN and the significance of errors. Psychophysiology, 42:151–160 (2005).
22. Harrington, D.L., Haaland, K.Y. Skill learning in the elderly: Diminished implicit and explicit memory for a motor sequence. Psychology and Aging, 7:425–432 (1992).

23. Hedden, T., Gabrieli, J.D.E. Insights into the aging mind: A view from cognitive neuroscience. Nature Reviews Neuroscience, 5:87–97 (2004).
24. Holroyd, C.B. A note on the oddball N200 and the feedback ERN. In M. Ullsperger, M. Falkenstein (Eds.), Errors, Conflict and the Brain. Current Opinions on Performance Monitoring (pp. 211–218). Dresden: Sächsisches Digitaldruck Zentrum GmbH (2003).
25. Holroyd, C.B., Coles, M.G. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychological Review, 109:679–709 (2002).
26. Howard, D.V., Howard, J.H., Jr. Age differences in learning serial patterns: Direct versus indirect measures. Psychology and Aging, 4:357–364 (1989).
27. Howard, D.V., Howard, J.H., Jr. Adult age differences in the rate of learning serial patterns: Evidence from direct and indirect tests. Psychology and Aging, 7:232–241 (1992).
28. Kliegl, R., Mayr, U., Krampe, R.Th. Time-accuracy functions for determining process and person differences: An application to cognitive aging. Cognitive Psychology, 26:134–164 (1994).
29. Kopp, B., Wolff, M. Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. Biological Psychology, 51:223–246 (2000).
30. Kray, J., Lindenberger, U. Adult age differences in task switching. Psychology and Aging, 15:126–147 (2000).
31. Kray, J., Lindenberger, U. Fluide intelligenz. In J. Brandtstädter, U. Lindenberger (Hrsg.), Entwicklungspsychologie der Lebensspanne. Ein Lehrbuch (S. 194–220). Stuttgart, Germany: Kohlhammer Verlag (2007).
32. Lindenberger, U., Kray, J. Kognitive Entwicklung. In S.-H. Filipp, U.M. Staudinger (Eds.), Entwicklungspsychologie des mittleren und höheren Erwachsenenalters (pp. 299–341). Göttingen, Germany: Hogrefe Verlag (2005).
33. Miltner, W.H.R., Braun, C.H., Coles, M.G.H. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. Journal of Cognitive Neuroscience, 9:788–798 (1997).
34. Nessler, D., Mecklinger, A. ERP correlates of true and false recognition after different retention delays: Stimulus- and response-related processes. Psychophysiology, 40:146–159 (2003).
35. Nieuwenhuis, S., Ridderinkhof, K.R., Talsma, D., Coles, M.G.H., Holroyd, C.B., Kok, A., van der Molen, M.W. A computational account of altered error processing in older age: Dopamine and the error-related negativity. Cognitive, Affective, & Behavioral Neuroscience, 2(1):19–36 (2002).
36. Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., Ridderinkhof, K.R. Electrophysiological correlates of anterior cingulated function in a go/no-go task: Effects of response conflict and trial type frequency. Cognitive, Affective, & Behavioral Neuroscience, 3:17–26 (2003).
37. Nissen, M.J., Bullemer, P. Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19:1–32 (1987).
38. Reber, A.S. Implicit learning and tacit knowledge. Journal of Experimental Psychology: General, 118:219–235 (1989).
39. Rüsseler, J., Hennighausen, E., Münte, T.F., Rösler, F. Differences in incidental and intentional learning of sensorimotor sequences as revealed by event-related brain potentials. Cognitive Brain Research, 15:116–126 (2003).
40. Rüsseler, J., Kuhlicke, D., Münte, T.F. Human error monitoring during implicit and explicit learning of a sensorimotor sequence. Neuroscience Research, 47:233–240 (2003).
41. Schultz, W. Getting formal with dopamine and reward. Neuron, 36:241–263 (2002).
42. Schultz, W., Dayan, P., Montague, P.R. A neural substrate of prediction and reward. Science, 275:1593–1599 (1997).
43. Seger, C.A. Implicit learning. Psychological Bulletin, 115:163–196 (1994).
44. Thorndike, E.L. Laws and hypotheses for behavior. In E.L. Thorndike (Ed.), Animal Intelligence (pp. 241–281). Darien, CT: Hafner Publishing Co. (1970) (Original work published 1911).
45. Turner, M.L., Engle, R.W. Is working memory capacity task dependent? Journal of Memory and Language, 28:127–154 (1989).

46. Ullsperger, M., von Cramon, D.Y. Subprocesses of performance monitoring: A dissociation of error processing and response competition revealed by event-related fMRI and ERPs. Neuroimage, 14:1387–1401 (2001).
47. Volkow, N.D., Logan, J., Fowler, J.S., Wang, G.-J., Gur, R.C., Wong, C., Felder, C., Gatley, S.J., Ding, Y.-S., Hitzemann, R., Pappas, N. Association between age-related decline in brain dopamine activity and impairment in frontal and cingulated metabolism. American Journal of Psychiatry, 157:75–80 (2000).
48. Wechsler, W. Handanweisung zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE) [Manual for the Hamburg-Wechsler intelligence test for adults]. Bern, Switzerland: Huber (1982).

An ERP-Approach to Study Age Differences in Cognitive Control Processes

Jutta Kray and Ben Eppinger

1 Introduction

The flexible adaptation to changes in the environment is one important feature of intelligent behaviour and is associated with the ability to efficiently control one's own processing. Cognitive control refers to the ability to guide thoughts and actions in accord with internal task goals. Controlling one's own behaviour is particularly required in situations that involve the flexible switching between multiple tasks, the selection and maintenance of task-relevant and the inhibition of task-irrelevant information, as well as the monitoring of error and conflict information (e.g. [22, 42, 47]). In this project phase we have focused on the investigation of age-related resource limitations in task switching and the inhibition of task-irrelevant information. Research in the later project phase was concerned with control processes related to the monitoring of error and conflict information (see Ferdinand et al., Error-Induced Learning as a Resource-Adaptive Process in Young and Elderly Individuals of this volume).

In this chapter, we will first review age-related resource limitations in control processes required for efficiently switching between tasks and for successfully inhibiting well-learned response tendencies. Then, we will briefly describe the paradigm that we applied to measure interactions between both types of control processes. The specific goal of our project was to identify the electrophysiological correlates of control processes during task preparation and task execution. By this we hope to gain a more fine-tuned analysis of age-related differences in cognitive control processing. In the result section, we will report the most important findings of two experiments on age-related differences in control processing. Finally, we will discuss these findings in the context of the actual control models and recent findings on age deficits in control processing.

J. Kray (✉)
Developmental Psychology Unit, Department of Psychology, Saarland University, Saarbrücken, Germany
e-mail: j.kray@mx.uni-saarland.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_5,

2 Cognitive Flexibility and Inhibition Limitations in Older Adults

2.1 Age Differences in Task Switching

One prototypical paradigm that has been used to measure the ability of the cognitive system to flexibly adapt to changing environments is the so-called task-switching paradigm (for reviews, see [33, 43]). At least one reason for its popularity is that the paradigm allows researchers to study multiple separate control processes such as task preparation, interference and switching processes within the same paradigm.

In task-switching procedures, participants are usually instructed to switch between two tasks A and B. For instance, in the one task (task A), participants are asked to respond to the meaning of words, and in the other task (task B), they are asked to respond to the colour in which the words are printed. Costs of switching between tasks can be determined by comparing performance in blocks of trials in which subjects switch between both tasks A and B (mixed-task blocks) with performance in blocks of trials in which they perform only one task A or B repeatedly (single-task blocks). A number of studies have demonstrated that reaction times (RT) are longer and error rates are larger in mixed-task blocks than in single-task blocks (e.g. [1, 2, 25]). This type of switching costs is henceforth termed general switch costs (cf. [29]) and is associated with "to be in a switch situation". General switching costs are thought to reflect control processes that are required for maintaining two task sets and for flexibly selecting between them. Costs of switching between task sets can also be determined within mixed-task blocks by comparing the performance on trials in which the task was changed (AB, BA) with the performance on trials in which the task was repeated (AA, BB). This type of costs has been termed switch costs (e.g. [48]) or specific switch costs (cf. [29]) and is thought to tap the switching process per se. So far a large variety of studies found that RTs are longer and error rates are higher in switch trials than in non-switch trials (for a review, see [43]).

There are now a number of studies that examined processing limitations of task-switching abilities in older age (for a review, see [24]; for a meta-analysis, see [55]). For theories of cognitive aging it is important to know whether age-related processing limitations occur in all kinds of cognitive processes, supporting the idea of general limitations in the speed of processing (cf. general slowing account of cognitive aging; see [31, 50]), or only in some cognitive processes, indicating process-specific limitations in the elderly. Empirical evidence so far suggests that older adults have mainly difficulties in task switching that are associated with maintaining task sets and selecting between them, that is, they show larger general switch costs than younger adults. In contrast, older adults have less difficulties in switching from one task to the other, that is, the magnitude of specific switch costs is often found to be similar in younger and older adults [7, 10, 26–29, 37, 46]. Hence, the good news is that not all cognitive control processes are limited in their functioning by old age, favouring the view of process-specific limitations of task-switching abilities in older adults.

This later view is also supported by studies using functional and structural imaging methods. For instance, Braver and colleagues [5] found that both types of switching costs were associated with neural activations in different brain regions. The magnitude of general switch costs is highly correlated with brain activity in the right anterior prefrontal cortex and the magnitude of the specific switch costs is highly correlated with brain activity in the left superior parietal cortex. This is interesting since data from functional as well as structural imaging studies suggest that age-related neural changes are most pronounced in the prefrontal cortex compared to other cortices (for reviews, see [20, 44, 45]), which is in line with the observed pattern of age differences in task switching.

2.2 Age Differences in Interference Control

Another well-known task to measure the efficiency of control processing is the Stroop paradigm [53]. This task requires maintaining a less-practised action intention (e.g. naming colours) and inhibiting a well-practised action tendency (reading words). Usually participants are presented words (i.e. red, blue, yellow and green) that are either printed in a compatible colour (the word "red" printed in "red") or in an incompatible colour (the word "red" printed in "blue") and they are instructed to name the colours. On incompatible trials the less-practised action intention (colour naming) interferes with the well-practiced action tendency of reading words. Therefore, larger RT and higher error rates can be observed on incompatible than on compatible trials, termed Stroop interference effect[1] in the following (for reviews, [35, 36]). The efficiency of the cognitive system to inhibit unwanted action tendencies is reflected in small interference effects.

Empirical evidence for age-related changes in interference control is not quite clear. A number of studies found greater Stroop interference effects for older than for younger adults [9, 12, 13, 28, 30, 52, 61], whereas a meta-analysis by Verhaeghen and De Meersman [56] did not support the view of process-specific age deficits in inhibitory processing (see also [51]). That is, larger Stroop interference effects in the elderly can also be explained by age differences in general slowing of processing speed. Thus, a further goal of this study was to contribute to this debate by specifying conditions under which age-related differences in interference control are more likely to occur.

To do this, we also varied the number of incompatible trials within a given block of trials (cf. [21, 54]). Indeed, a number of studies have shown that demands on interference control are increased when incompatible trials are less frequent. Put differently, if the cognitive system adapts to easier task conditions (high number of compatible trials within a block) in which no interference occurs and is then confronted with an incompatible trial, more resources are required to inhibit the

[1] Note that there are different ways to determine the interference effect, depending on which type of baseline condition is used for comparison.

unwanted action tendency. As a consequence, the Stroop interference effect is larger. Conversely, if the cognitive system adapts to situations in which demands on interference control are high, the Stroop interference effect is smaller [21, 54].

2.3 Interactions Between Cognitive Control Processes and Their Temporal Dynamics

The general goal of this project was to assess age-related resource limitations in adaptive behaviour, in particular, in the ability to efficiently switch between tasks and to inhibit well-learned action tendencies. Our specific goal was to examine interactions between these cognitive control processes as well as to determine their temporal dynamics. To investigate interactions between control processes, we combined the above described well-known Stroop stimuli with a cue-based task-switching paradigm. Thus, the participants in our studies were required to switch between two tasks, that is, they either had to respond to the colours (task A) or to the words (task B). Demands on cognitive control processes are largest when subjects have to switch between tasks and to inhibit the currently irrelevant action tendency. Which task they should perform in a given trial was indicated by a task cue (FAR for naming the colour (in German "Farbe") and WOR for reading the word (in German "Wort")). Subsequently, the Stroop stimulus was presented (for details, see Sect. 3). This type of paradigm has several advantages. First, it allows separating cognitive activity recruited for preparing the next task from cognitive activity recruited for executing the actual tasks. Second, it also allows separating cognitive activity for task preparation from cognitive activity required for disengaging from the previous task sets (cf. [39]).

Therefore, this type of paradigm is well suited to examine whether different kinds of brain regions are recruited for task preparation and task execution by using functional imaging methods. Indeed, MacDonald et al. [34] demonstrated a dissociation between two cognitive control functions. During the task-preparation interval (between task cue and target presentation), they found larger neural activity in the left dorsolateral prefrontal cortex (DLPFC) when subjects prepared for the less practised colour naming than for the well-practised word reading tasks. During the task-execution interval (between target presentation and the next task cue), they observed a larger neural activity in the anterior cingulate cortex (ACC) on incompatible than on compatible Stroop stimuli. On the basis of these results and other findings, it has been concluded that the DLPFC plays an important role for the active representation and maintenance of task instructions or S-R rules (e.g. [41]), while the ACC is involved in the monitoring and evaluation of conflicts (e.g. [4, 6, 36]). In general, current models of cognitive control suggest that the control of goal-directed behaviour is implemented in the brain by a distributed network with anatomically dissociable subcomponents. According to the network view, the DLPFC biases task-appropriate behaviour and the ACC corresponds to an evaluation system that indicates when more control is needed [8, 19, 23]. Because of the high

spatial resolution of fMRI, this method is well suited to determine what kinds of brain regions are recruited for adapting to increased control demands.

The main goal of our project was to examine the temporal course of neural activity during task preparation and task execution when the demands on cognitive control are increased. An event-related potential (ERP) approach is particularly useful in this respect, because its temporal resolution is in the millisecond range, which allows the measurement of the time course of neuronal activity with high precision. In the first experiment, we aimed at identifying ERP-correlates of age-specific resource limitations in task preparation and inhibition processes. In the second experiment, we additionally manipulated the frequency of incompatible Stroop trials to investigate how flexible the cognitive system is in allocating control resources.

3 Methods Applied

3.1 Participants

Fourteen younger adults (mean age = 21.7 years, SD = 2.15, 6 females) and 14 older adults (mean age = 62.9 years, SD = 1.9, 6 females) participated in the first experiment. All participants indicated themselves to be healthy, having a right-hand preference, no colour blindness and no history of neurological or psychiatric problems. As often reported in the aging literature [50], we obtained no reliable age differences in a subtest of semantical knowledge, but significant differences in perceptual speed of processing, indicating the representativity of the younger and older subsample (for details, see [28]).

A subsample of participants of the first study also participated in the second experiment, 12 younger adults (mean age = 21.3 years, SD = 1.8, 6 female) and 12 older adults (mean age = 63.7 years, SD = 2.6, 6 female) (for further details, see [13]).

3.2 Procedure

The participants were presented four words (i.e. RED, BLUE, YELLOW and GREEN) whereas the display colours were either compatible or incompatible with the word meaning. The participants were instructed to perform two tasks, which will be referred to as “colour task” and “word task” in the following. In the word task, participants responded to the meaning of the word, and in the colour task to the display colour with one of the same four response keys.

In single-task conditions, the participants performed only the colour or the word task, and in mixed-task conditions, they were instructed to switch between both tasks. Which task to perform next was indicated by a task-set cue (the German letter string -wor- for the word task and -far- for the colour task).

The trial procedure was identical for single- and mixed-task blocks. Each trial started with a task-set cue (i.e. -far- or -wor-) that was presented for 500 ms, followed by a blank screen of 1,800 ms (see Fig. 1). Before the target (the Stroop word), a fixation cross was displayed for 200 ms. Then, the target stimulus was presented for 300 ms, followed by a blank screen until the response was made. ERPs were recorded in a task-cue interval, in which the subjects prepared for the upcoming task, and in a target interval, in which responses to Stroop stimuli had to be given.

Fig. 1 The trial procedure

Stimuli, tasks and trial procedure were identical in the first and the second experiment. In the first experiment, the number of incompatible trials in a block was 50%. In the second experiment, we additionally manipulated the frequency of incompatible trials within a block that was either high (80% of the trials) or low (20% of the trials).

3.3 Data Recordings

We used an IBM compatible computer for registration of reaction times (RTs) and errors and a CTX 17-inch colour monitor with a black background for stimulus presentation. Responses were registered using external response buttons.

EEG and EOG activities were recorded continuously (Neuroscan Synamps and Scan 4.2 acquisition software) from 64 tin electrodes (10–10 system) using an elastic cap (Electrocap International). We used the left mastoid as reference and the right mastoid as an active channel. The EEG and EOG signals were online bandpass filtered (DC – 70 Hz, 50 Hz notch filter) and digitized at 500 Hz. Impedances were kept below 5 kΩ (for details, see [13, 28]).

4 Results

In this section, we will report the most important findings of two experiments on age-related changes in cognitive control processes and discuss them in the context of previous findings. At first we report the behavioural results and then the ERP results.

4.1 Behavioural Results

Consistent with a number of previous findings (e.g. [25, 29, 37]; for a review, see [24]; for a meta-analysis, see [55]), we found an age-related increase in general

switch costs and no age differences in specific switch costs. We also found that age differences in general switch costs were more pronounced in the word reading than colour naming task (see Fig. 2). This finding is consistent with others that demonstrated that the magnitude of switch costs depends on the dominance of a well-practised task relative to a less-practised task. For instance, Allport et al. [1], using Stroop stimuli in a task-switching paradigm, found that switching to the well-learned task (here the word naming task) results in greater switching costs. Thus, it appears that older adults have more problems to efficiently switch between tasks when word processing needs to be inhibited in the previous trial than younger adults do.

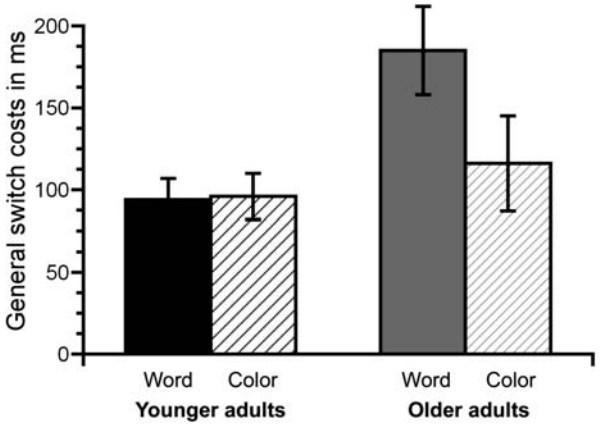


Fig. 2 General switch costs (latencies in ms) as a function of age group (younger and older adults) and task (word and colour task). Error bars refer to the standard error of the mean (SE)

As noted earlier, we were specifically interested in the interaction among cognitive control processes (see Sect. 2.3). Interestingly, our results showed that the Stroop interference effect was larger when subjects had to switch between tasks compared to when subjects were required to perform only one task in a block (e.g. [59]). Thus, subjects are less efficient in interference control when the demands on cognitive control are increased, that is, when participants have to switch between tasks. However, we did not find interactions with age.

Also in line with findings of a meta-analysis [56], results of our first study indicated that age differences in the Stroop interference disappeared when age differences in general speed were controlled.[2] This finding suggests that there are no age-specific resource limitations in interference control. However, age differences in interference control might occur when demands in control processing are increased. Therefore, we increased the demands on interference control by varying

[2] Age differences in interference effects were no longer significant on the basis of proportional scores (incompatible/compatible), which take age differences in baseline performance (here: latencies in compatible trials) into account, only on the basis of difference scores (incompatible – compatible).

the frequency of incompatible trials in our second study. Consistent with previous studies, we found larger interference effects when the frequency of conflict trials was low (see Fig. 3; cf. [21, 54]). Hence, subjects were less adapted to conflict when the frequency of incompatible trials was low and had to engage more in conflict monitoring when an infrequent incompatible trial occurred. Of most relevance for the second study was the fact that age differences in the Stroop interference effect were influenced by the frequency manipulation. Older adults showed larger interference effects than younger adults when the frequency of conflict trials was low (see Fig. 3). This effect was only found for the colour naming task, but not for the word reading task, and was significant even after controlling for the effects of general slowing, that is, on the basis of proportional scores. Hence, there is evidence for age differences in the Stroop interference effect, but it is limited to situations of high demands on interference control.

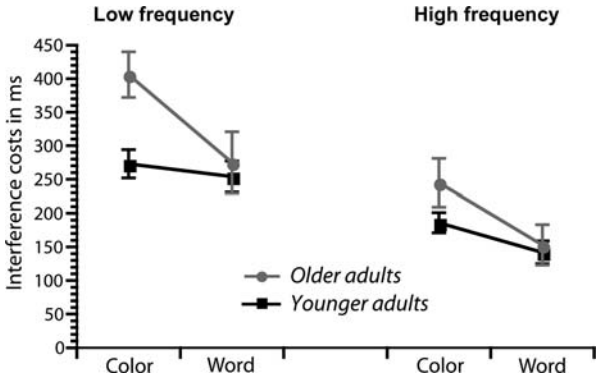


Fig. 3 Interference costs (latencies in ms) as a function of age group (younger and older adults) and task (word and colour task) and low (*left side*) and high frequency of incompatible trials (*right side*). Error bars refer to the standard error of the mean (SE)

4.2 ERP-Results

ERP data will be reported in two sections. In the first section we report analyses of neural activity that is related to preparatory processes and the ERPs were averaged time-locked to the onset of task-cue presentation (see Fig. 1). In the second section we report analyses of neural activity that is related to target processing and response execution. Thus, the ERPs were averaged time-locked to target and response onset. We will primarily review results on age-related changes in ERP correlates of task preparation and inhibitory processes.

4.2.1 Age Differences in ERP Correlates of Task-Preparation Processes

The analysis of task-cue related neural activity indicated age-related changes in two ERP components: in the P300 and in a contingent negative slow wave (CNV). Figure 4 displays ERP grand averages in the task-cue interval at three central

electrodes elicited in single- and mixed-task blocks separately for younger and older adults aggregated across the colour naming and word naming task. For all conditions a large positive component, the P3, was evoked at parietal electrodes for younger and older adults. Towards the end of the task-cue interval, a CNV emerged at central recording sites for the mixed relative to single-task blocks in the older age group.

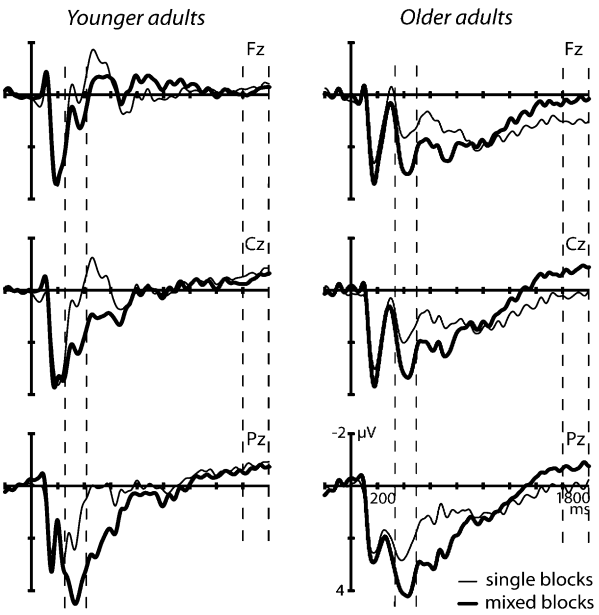


Fig. 4 Grand average ERPs in the cue interval for single (*thin line*) versus mixed-task (*thick line*) blocks separately for younger and older adults aggregated across tasks at the three midline electrodes (FZ, CZ and PZ). The *vertical bars* indicate cue onset, tick spacing in the *x*-axis is 200 ms, and the broken lines indicate the time windows that were used for statistical analysis of the P300 and CNV

The first noteworthy finding was a parietally distributed P300 was evoked by task cues that was larger under switching conditions (i.e. in mixed-task blocks) than non-switching conditions (i.e. in single-task blocks), which is consistent with other findings (e.g. [59]). Generally, the P300 is assumed to reflect processes that encode and update the currently relevant task context (e.g. [11, 38]). In the present study, the P300 presumably reflects the updating of task sets for the word or colour task. Age-related changes in this component are primarily related to the P300 latency that is significantly slowed in the older group and under switching conditions. Age-related slowing of P300 peak latency is a well-replicated finding in the area of cognitive aging; however, most studies focused on age differences in the implementation of attentional control as measured with the Oddball paradigm (for a review, see [49]).

In contrast to the age effects for P300 peak latency, we found no reliable age effect in the P300 amplitude. Most obvious was that the P300 amplitude was

substantially larger for switching conditions in both tasks. Although we found no age-related changes in the P300 amplitude, the topography of the P300 was significantly modulated by age but only in the word task. In the word task, in which general switch costs were also higher for older adults, the P300 topography took the form of a more flattened anterior–posterior distribution. Specifically, we found a loss of the centro-parietal focus in the older age group. Age-related changes in the topography of the P300 peak amplitude of similar kinds have been reported in other ERP studies on cognitive aging (e.g. [17]). The functional and neuroanatomical factors contributing to the modified P300 topography in the elderly are still a matter of debate. However, we suggest that the flattened P300 topography and the enhanced general switch costs in the word task reflect that older adults may recruit frontal areas to a larger extent for the more demanding implementation of the task set of the word task (cf. [40]).

A second important finding of our study is that we found age-related changes in the CNV. The CNV is assumed to be associated with the ability to maintain task-set representations over time. In the present study, no reliable CNV differences between single and mixed blocks were obtained for younger adults (see Fig. 4). In contrast, older adults showed a substantially larger CNV under switching conditions, suggesting that they were differentially engaged in the maintenance of task sets in mixed compared to single task blocks. If we take the CNV as an indicator of the ability to actively maintain task-set representations over time, then the findings suggest that older adults have problems in maintaining the currently relevant task set under mixed-task conditions (but see [59]; for an alternative interpretation, see [28]).

In our second study we investigated age differences in the flexibility of resource-adaptive behaviour by manipulating the degree of interference during task performance. It was expected that in blocks with a low frequency of incompatible trials, control demands on task preparation are decreased whereas demands on conflict monitoring are increased, and vice versa. However, age effects in the ERP correlates of task preparation did not interact with the frequency manipulation (for other effects of the frequency manipulation, see [13]).

4.2.2 Age Differences in ERP Correlates of Interference Control

The analyses of ERP components in the target interval allowed us to examine age-related differences in interference control during target processing and response execution. Findings from several recent ERP studies suggest that stimulus-driven inhibitory control is reflected in a negativity for incompatible Stroop trials (which will be termed Ni in the following). This component has been assumed to reflect early conflict detection (e.g. [32, 59, 60]). A second negativity occurs during response execution on incompatible Stroop trials, which is called correct response negativity (CRN).

Results of our first experiment showed that Ni latency was slower for the colour naming than the word reading task, which is consistent with a greater behavioural Stroop interference in the colour naming than in the word reading task (see Fig. 2).

Moreover, the Ni was substantially slowed for older than for younger adults. Age effects were even more pronounced in the colour naming task when subjects had to switch between tasks, thus when control demands were increased. Furthermore, consistent with a number of other findings (e.g. [32, 60]), we obtained a larger Ni amplitude for incompatible than for compatible trials. However, this effect was clearly present in both age groups and in both tasks, supporting the view that the Ni component reflects general conflict processing that is required whenever a target stimulus involves ambiguous information (e.g. [32, 60]). Thus, it appears that the Ni reflects a general mechanism of early conflict detection that is relatively invariant across tasks and age. Hence, we only found evidence that conflict detection in slowed in the elderly.

In contrast, results of both studies provided evidence for age-related differences in response-related conflict processing. For younger adults, a negativity (CRN) was obtained that was larger for incompatible than for compatible trials. As no such compatibility effect was found for older adults, this can be taken as evidence for age differences in response-related conflict processing. Hence, it appears that younger adults are better able than older adults to discriminate between incompatible and compatible trials at a response-related processing stage. A similar negative deflection, termed Error-related Negativity (ERN, [18]) or Error negativity (Ne, [14]), peaking around 80 ms after the response, is often observed in erroneous responses and has been considered as a part of a more general executive control system that monitors for conflicts and errors. The attenuation of the Ne in older adults [15, 16, 19] has been taken as evidence for a lower flexibility of error and action monitoring in the elderly. Even though more research is required to elucidate the functional processes reflected in the CRN, the present results of age differences in the CRN suggest, similar to the ERN/Ne, age-related changes of the action monitoring system.

Results of our second study replicated and extended this result, that is, we found a larger CRN for incompatible than for compatible trials only for the younger adults (see also [57, 58]). Furthermore, the compatibility effect in the CRN for younger adults was sensitive to the probability manipulation, i.e. it was larger when incompatible trials were less frequent (see Fig. 5). This pattern of results in the CRN amplitudes parallels the behavioural Stroop interference effect that was larger when demands on conflict processing were increased (see Fig. 3). This finding also suggests that younger adults are able to adjust their behaviour depending on the task context, which means that they are well prepared for conflict when incompatible trials are frequent and less prepared when conflict is decreased and incompatible trials are rare.

Our results are also consistent with recent findings from a study of Bartholow and colleagues [3], who manipulated the frequency of incompatible trials in the Eriksen Flanker task. They also found that the CRN was sensitive not only to response conflict, but also to conflict that emerges, when expectancies about the compatibility of the target stimulus were violated.

In contrast to younger adults, older adults did not show reliable differences in the CRN between conditions, with enhanced amplitudes, independently of the

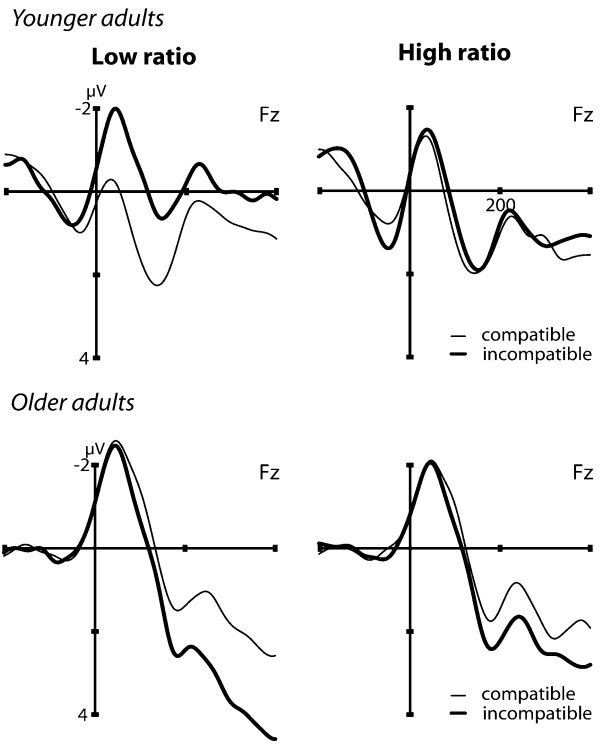


Fig. 5 Grand average ERPs in the response interval for compatible (*thin line*) vs. incompatible (*thick line*) correct trials separately for younger and older adults as a function of low versus high frequency (ratio) of incompatible trials at FZ. The *vertical bars* indicate response onset, and tick spacing in the *x*-axis is 200 ms

frequency manipulation. Thus, older adults were not able to flexibly adapt to changes of conflict demands. The finding that the CRN did not differentiate between compatible and incompatible trials in the older age group confirms the findings of our first study and suggests impairments of older adults in discriminating compatible from incompatible trials. Our data are also consistent with findings from Gehring and Knight [19]. They found that CRNs were enhanced for patients with frontal lobe lesions. Gehring and Knight [19] suggested that these patients might suffer from an impaired representation of the contextually appropriate stimulus response mapping, and by this, are not able to distinguish between what was a correct response and what was not. Applying this argumentation to older adults, it is conceivable that in highly demanding task situations older adults suffer from a compromised representation of the actually relevant task set (and thus the correct response) even on compatible trials, leading to conflict when the actual response is matched with the impaired representation. As a consequence, they are not able to build up expectancies about the compatibility of the target stimulus, which might explain the absence of a frequency effect in the older age group.

5 Summary and Conclusions

In sum, the results of our studies have implications on current theories of cognitive aging and provide a number of new theoretical insights in the field of age-related resource limitations of adaptive behaviour.

First, the behavioural data of the two studies suggest, in line with previous findings [29, 37], that age-related resource limitations primarily occur when older adults have to select and maintain task-relevant information, thus at a general level of switching between task sets. In contrast, older adults have no problems when they have to reconfigure tasks on a trial-to-trial basis, thus at a specific level of task switching. Moreover, in the first study we did not find evidence for age-related impairments in the ability to inhibit irrelevant action tendencies, which is also in line with previous findings [56]. The new behavioural findings of our studies are that older adults have even greater difficulties in task-set selection when they are switching to a well-learned activity (the reading task) that was strongly inhibited in the previous trial. Moreover, older adults have more difficulties in inhibitory processing than younger adult do, when demands on interference control are increased (i.e. when incompatible trials are less frequent and therefore less expected).

Second, we identified two ERP components that varied with switching demands during task preparation, the P300 and the CNV. The P300 is thought to reflect processes related to encoding and updating of task-relevant information (e.g. [11, 38]). The P300 latency was increased in the elderly, especially in mixed-task blocks, indicating that older adults are slowed in the updating of task-relevant information when they have to switch between tasks. The P300 amplitude elicited by the task-set cue was larger for mixed-task blocks than for single-task blocks with a parietal maximum in the younger group and a broader distribution in the older group in the word task. This finding suggests age-related deficits in updating of currently relevant tasks, primarily when switching from a less-practised task is required. Furthermore, only older adults showed an enhanced CNV under switching conditions, suggesting that older adults have problems to maintain task-relevant information over time.

Third, results of our studies also identified two ERP components that varied with inhibitory demands, an early negativity related to the processing of conflicting task information (termed Ni; cf. [32, 60]), and a late negativity related to response-related processing (termed CRN; see [57, 58]). The Ni is thought to reflect conflict detection when the target stimulus contains ambiguous information. The latency of this component was substantially increased for the colour naming than the word reading task under switching conditions and this difference varied with age. Thus, older adults were much slower than younger adults in conflict detection when demands on cognitive control were increased. Although the Ni was larger for incompatible than for compatible trials [32, 60], the Ni amplitude did not vary across tasks and age, suggesting no qualitative changes in conflict processing in the elderly and for different task domains. Furthermore, the response-related negativity on correct trials (CRN) was found to be larger for incompatible than compatible Stroop stimuli. However, this was only the case for the younger age group, whereas the CRN occurred on both types of trials in the older group. Moreover, the compatibility effect in the

CRN varied as a function of conflict ratio in younger but not in older adults, which points to the view that older adults are impaired in the flexible adaptation to changing demands on conflict processing. As the CRN has been considered to reflect the efficiency of the conflict monitoring system [3, 19], results of our studies suggest that the older adults were less efficient in discriminating between incompatible and compatible trials at a response-related processing stage and by this showed resource limitations in action monitoring as well as in flexibly adapting to changes of conflict demands.

Taken together, our findings are inconsistent with the view that cognitive aging is mediated by a single, global mechanism that affects all type of cognitive processes such as a general age-related decline in speed of information processing. Instead, our findings support the view of age-specific resource limitations in cognitive control processing. On the behavioural level, we found age-related resource limitations in maintaining and selecting between task sets and in inhibitory processing when the demands on cognitive control were high. This view was further confirmed by age-related differences in ERP correlates of task-preparation and inhibitory processing. Age-related slowing was observed for task-set updating and conflict detection. Age-specific processing limitations were obtained during task-set maintenance and response-related conflict processing, as well as in the flexibility adaptation to conflict demands. Generally, the ERP-approach seems to be a useful tool for obtaining a detailed view on age differences in cognitive control processing.

Acknowledgments This research was supported by the Deutsche Forschungsgemeinschaft (grant SFB 378, EM 2). We wish to thank Oliver John and Barbara Mock for their support during data collection and Axel Mecklinger for very helpful comments.

References

1. Allport, D.A., Styles, E.A., Hsieh, S. Shifting intentional set: Exploring the dynamic control of tasks. In C. Emilia, M. Moscovitch (Eds.), Attention and Performance XV (pp. 421–452). Cambridge, MA: The MIT Press (1994).
2. Allport, D.A., Wylie, G. Task-switching, stimulus-response bindings, and negative priming. In S. Monsell, J. Driver (Eds.), Control of Cognitive Processes: Attention and Performance XVIII. Cambridge, MA: MIT Press (1999).
3. Bartholow, B.D., Pearson, M.A., Dickter, C.L., Sher, K.J., Fabiani, M., Gratton, G. Strategic control and medial frontal negativity: Beyond errors and response conflict. Psychophysiology, 42:33–42 (2005).
4. Botvinick, M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D. Conflict monitoring and cognitive control. Psychological Review, 108:624–652 (2001).
5. Braver, T.S., Reynolds, J.R., Donaldson, D.I. Neural mechanisms of transient and sustained cognitive control during task switching. Neuron, 39:713–726 (2003).
6. Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., Cohen, J.D. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science, 280:747–749 (1998).
7. Cepeda, N.J., Kramer, A.F., Gonzales De Sather, J.C.M. Changes in executive control across the life span: Examination of task-switching performance. Developmental Psychology, 37:715–730 (2001).

8. Cohen, J.D., Botvinick, M., Carter, C.S. Anterior cingulate and prefrontal cortex: Who's in control? Nature Neuroscience, 3:421–423 (2000).
9. Comalli, P.E.J., Wapner, S., Werner, H. Interference effects of Stroop color-word test in childhood, adulthood, and aging. Journal of Genetic Psychology, 100:47–53 (1962).
10. DiGirolamo, G.J., Kramer, A.F., Barad, V., Cepeda, N.J., Weissman, D.H., Milham, M.P., et al. General and task-specific frontal lobe recruitment in older adults during executive processes: A fmri investigation of task-switching. Neuroreport, 12(9):2065–2071 (2001).
11. Donchin, E., Coles, M.G. Is the P300 component a manifestation of context updating? Behavioral and Brain Sciences, 11:357–427 (1988).
12. Dulaney, C.L., Rogers, W.A. Mechanisms underlying reduction in Stroop interference with practice for young and old adults. Journal of Experimental Psychology: Learning, Memory and Cognition, 20:470–484 (1994).
13. Eppinger, B., Kray, J., Mecklinger, A., John, O. Age differences in task switching and response monitoring: Evidence from ERPs. Biological Psychology, 75:52–67 (2007).
14. Falkenstein, M., Hohnsbein, J., Hoormann, J., Blanke, L. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In C.H.M. Brunia, A.W.K. Gaillard, A. Kok (Eds.), Psychophysiological Brain Research (pp. 192–195). Tilburg: Tilburg University Press (1990).
15. Falkenstein, M., Hoormann, J., Hohnsbein, J. Changes of error-related ERPs with age. Experimental Brain Research, 138:258–262 (2001).
16. Falkenstein, M., Hoormann, J., Hohnsbein, J. Inhibition-related ERP components: Variation with modality, age, and time-on-task. Journal of Psychophysiology, 16:167–175 (2002).
17. Friedman, D., Kazmerski, V., Fabiani, M. An overview of age-related changes in the scalp distribution of P3b. Electroencephalography and Clinical Neurophysiology, 104:498–523 (1997).
18. Gehring, J.W., Goss, B., Coles, M.G., Meyer, D.E., Donchin, E. A neural system for error detection and compensation. Psychological Science, 4:385–390 (1993).
19. Gehring, J.W., Knight, R.T. Prefrontal cingulate interactions in action monitoring. Nature Neuroscience, 3:516–520 (2000).
20. Hedden, T., Gabrieli, J.D.E. Insights into the ageing mind: A view from cognitive neuroscience. Nature Neuroscience Reviews, 5:87–96 (2004).
21. Jacoby, L.L., Lindsay, D.S., Hessels, S. Item-specific control of automatic processes: Stroop process dissociations. Psychonomic Bulletin and Review, 10:638–644 (2003).
22. Jonides, J., Smith, E.F. The architecture of working memory. In E.D. Rugg (Ed.), Cognitive Neuroscience. Studies in Cognition. Cambridge, MA: MIT Press (1997).
23. Kerns, J.G., Cohen, J.D., Mac Donald, A.W., Cho, R.Y., Stenger, V.A., Carter, C.S. Anterior cingulate conflict monitoring and adjustments in control. Science, 303:1023–1026 (2004).
24. Kramer, A.F., Kray, J. Aging and divided attention. In E. Bialystok, F.I.M. Craik (Eds.), Lifespan Cognition: Mechanisms of Change. New York: Oxford University Press (2006).
25. Kray, J. Task-set switching under cue-based and memory-based switching conditions in younger and older adults. Brain Research, 1105:83–92 (2006).
26. Kray, J., Eber, J., Karbach, J. Verbal self-instructions in task switching: A compensatory tool for action-control deficits in childhood and old age? To appear in: Developmental Science, 11:223–236 (2008).
27. Kray, J., Eber, J., Lindenberger, U. Age differences in executive functioning across the lifespan: The role of verbalization in task-preparation. Acta Psychologica, 115:143–165 (2004).
28. Kray, J., Eppinger, B., Mecklinger, A. Age differences in attentional control: An event-related potential approach. Psychophysiology, 42:407–416 (2005).
29. Kray, J., Lindenberger, U. Adult age differences in task switching. Psychology and Aging, 15:126–147 (2000).
30. Li, K.Z.H., Bosman, E.A. Age differences in Stroop-like interference as a function of semantic relatedness. Aging, Neuropsychology, and Cognition, 3:272–284 (1996).
31. Lindenberger, U., Mayr, U., Kliegl, R. Speed and intelligence in old age. Psychology and Aging, 8:207–220 (1993).

32. Liotti, M., Woldorff, M.G., Perez III., R., Mayberg, H.S. An ERP study of the temporal course of the Stroop color-word interference effect. Neuropsychologia, 38:701–711 (2000).
33. Logan, G.D. Executive control of thought and action: In search of the wild homunculus. Psycholological Science, 12:45–48 (2003).
34. MacDonald III, A.W., Cohen, J.D., Stenger, V.A., Carter, C.S. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science, 288:1835–1838 (2000).
35. MacLeod, C.M. The stroop task: The 'gold standard' of attentional measures. Journal of Experimental Psychology: General, 121:12–14 (1992).
36. MacLeod, C.M., MacDonald, P.A. Interdimensional interference in the Stroop effect uncovering the cognitive and neural anatomy of attention. Trends in Cognitive Sciences, 4:383–391 (2000).
37. Mayr, U. Age differences in the selection of mental sets: The role of inhibition, stimulus ambiguity, and response-set overlap. Psychology and Aging, 16:96–109 (2001).
38. Mecklinger, A., Ullsperger, P. P3 varies with stimulus categorization rather than probability. Electroencephalography and Clinical Neurophysiology, 86:395–407 (1993).
39. Meiran, N. Reconfiguration of processing mode prior to task performance. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22:1423–1442 (1996).
40. Milham, M.P., Erickson, K.L., Banich, M.T., Kramer, A.F., Webb, A., Wszalek, T., et al. Attentional control in the aging brain: Insights from an fMRI study of the Stroop task. Brain and Cognition, 49:277–296 (2002).
41. Miller, E.K., Cohen, J.D. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24:167–202 (2001).
42. Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T.D. The unity and diversity of executive functions and their contributions to a complex "frontal lobe" task: A latent variable analysis. Cognitive Psychology, 41:49–100 (2000).
43. Monsell, S. Task switching. Trends in Cognitive Sciences, 7:134–140 (2003).
44. Raz, N. Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F.I.M. Craik, T.A. Salthouse (Eds.), Handbook of Aging and Cognition (2nd Ed., pp. 1–90). Mahwah, NJ: Lawrence Erlbaum (2000).
45. Raz, N. The aging brain observed in vivo: Differential changes and their modifiers. In R. Cabeza, L. Nyberg, D. Park (Eds.), Cognitive Neuroscience of Aging: Linking Cognitive and Cerebral Aging (p. 400). New York: Oxford University Press (2005).
46. Reimers, S., Maylor, E.A. Task switching across the lifespan: Effects of age on general and specific switch costs. Developmental Psychology, 41:661–671 (2005).
47. Ridderinkhof, K.R., van den Wildenberg, W.P.M., Segalowitz, S.J., Carter, C.S. Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. Brain and Cognition, 56:129–140 (2004).
48. Rogers, R.D., Monsell, S. Costs of a predictable switch between simple cognitive tasks. Journal of Experimental Psychology: General, 124:207–231 (1995).
49. Polich, J. Meta-analysis of P300 normative aging studies. Psychophysiology, 33:334–353 (1996).
50. Salthouse, T.A. The processing-speed theory of adult age differences in cognition. Psychological Review, 103:403–428 (1996).
51. Salthouse, T.A., Meinz, E.J. Aging, inhibition, working memory, and speed. Journals of Gerontology: Psychological Sciences, 50B:P297–P306 (1995).
52. Spieler, D.H., Balota, D.A., Faust, M.E. Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer type. Journal of Experimental Psychology: Human Perception and Performance, 22:461–479 (1996).
53. Stroop, J.R. Studies of interference in serial and verbal reactions. Journal of Experimental Psychology, 18:643–662 (1935).

54. Tzelgov, J., Henik, A., Berger, J. Controlling Stroop effects by manipulating expectations for color words. Memory and Cognition, 20:727–735 (1992).
55. Verhaeghen, P., Cerella, J. Aging, executive control, and attention: A review of metaanalyses. Neuroscience and Biobehavioral Reviews, 26(7):849–857 (2002).
56. Verhaeghen, P., De Meersman, L. Aging and the Stroop effect: A meta-analysis. Psychology and Aging, 13:120–126 (1998).
57. Vidal, F., Burle, B., Bonnet, M., Grapperon, J., Hasbroucq, T. Error negativity on correct trials: A reexamination of available data. Biological Psychology, 64:265–282 (2003).
58. Vidal, F., Hasbroucq, T., Grapperon, J., Bonnet, M. Is the 'error negativity' specific to errors? Biological Psychology, 51:109–128 (2000).
59. West, R. The effects of aging on controlled attention and conflict processing in the Stroop task. Journal of Cognitive Neuroscience, 16:103–113 (2004).
60. West, R., Alain, C. Age related decline in inhibitory control contributes to the increased Stroop effect observed in older adults. Psychophysiology, 37:179–189 (2000).
61. West, R., Baylis, G.C. Effect of increased response dominance and contextual disintegration on the Stroop interference effect in older adults. Psychology and Aging, 13:206–217 (1998).

Simulating Statistical Power in Latent Growth Curve Modeling: A Strategy for Evaluating Age-Based Changes in Cognitive Resources

Timo von Oertzen, Paolo Ghisletta, and Ulman Lindenberger

1 Introduction

Variability across and within individuals is a fundamental property of adult age changes in behavior [20, 21, 24]. Some people seem young for their age, others seem old; shining examples of older individuals who maintained high levels of intellectual functioning well into very old age, such as Johann Wolfgang von Goethe or Sophocles, stand in contrast to individuals whose cognitive resources are depleted by the time they reach later adulthood. A similar contrast exists between different intellectual abilities. For example, if one looks at the speed needed to identify and discriminate between different percepts, one is likely to find monotonic decline after late adolescence and early adulthood. But if one looks at verbal knowledge, one will find age stability or positive change into very old age [36]. As a general rule, tasks that assess individual differences in speed, reliability, and coordination of elementary processing operations show greater decline, whereas tasks that assess individual differences in acquired knowledge show less decline.

The simultaneous presence of resource growth, maintenance, and decline, both across individuals and across abilities, calls for statistical methods that are able to efficiently capture both the commonalities and the differences of age-based changes in levels of functioning across the lifespan [5]. In this context, a family of methods known as latent growth curve models (LGCMs), multi-level models, random-coefficient modeling has gained prominence in recent years [15]. Despite their widespread and increasing application, central statistical properties of these models have not yet been explored or formally analyzed. In this chapter, we introduce a general strategy for evaluating the suitability of LGCM for charting lifespan changes in behavior, with a specific emphasis on statistical power.

LGCMs [26] are a particular set of structural equation models (SEMs) aimed at describing the general, average trend in change as well as the individual differences around the group trend. Extensions allow for including predictors of interindividual

T. von Oertzen (✉)
Center for Lifespan Psychology, Max Planck Institute of Human Development, Berlin, Germany
e-mail: vonoertzen@mpib-berlin.mpg.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_6,

differences in change parameters obtained from the time series analyzed. Because longitudinal data are composed of repeated measures, the non-zero within-person correlation violates the common statistical assumption of independence in ordinary least squares regression analysis. A family of analyses for correlated data (e.g., multi-level, hierarchical linear, mixed effects, and random effects models) thus provides an appealing analytical strategy for longitudinal data [18]. The LGCM and the correlated data approach to repeated measures are statistically equivalent [30, 33]. For simplicity, we will here subsume both statistical approaches under the heading of LGCM.

LGCMs have become the favorite analytical tool of many psychological researchers for theoretical investigations about development and change phenomena. Several long-standing goals of longitudinal analyses may be achieved by implementing proper LGCMs under specific assumptions [4]. Because of their popularity, many software packages allow for easily reproducible LGCM analyses, either within the more general SEM framework or within the analogous correlated data approach (software implementation may highlight and optimize different statistical aspects; see [13, 22]). However, not all scientific inquiries nor all longitudinal data are amenable to LGCM analyses, and some warnings have been raised as to the limits of LGCMs [25, 34]. In particular, LGCMs have been utilized to reach substantive conclusions about concomitant interindividual differences in a multivariate space, especially in cognitive aging research. Yet, in samples of aging individuals it is often very hard to detect interindividual differences in change and subsequently, covariances among change components. Given the need to more fully understand statistical properties of LGCMs simulation work to this end has appeared in the literature (e.g., [11, 14, 16]).

LGCMs are most commonly computed with SEM software by applying a maximum likelihood estimation procedure to a moment matrix containing covariance and mean information about the repeated measures. For the common case of incomplete data, the Full Information Maximum Likelihood (FIML) variant allows analyzing raw data of all observations, without excluding observations with an incomplete data vector (cf. [1, 19, 27]). The FIML algorithm is now the choice of incomplete data treatment in many SEM software packages, including Mx [32], Lisrel [17], AMOS [2], EQS [6], and MPlus [31]. The mathematical formulation of the FIML algorithm can be found in the original source by Lange et al. (1976) or in some SEM manuals cited above. However, the general implementation of Lange et al.'s formulation within each SEM software package and the remaining elements of the general computation procedure used in the parameter estimation process are not easily documented, hence generally not available to SEM users.

In this chapter, we aim at (a) presenting a general simulation procedure for testing specific statistical properties of LGCMs and (b) describing the mathematical formulation of the estimation procedure adopted within our data-generation-plus-analysis engine. The simulation tool can be found at `http://www.mpib-berlin.mpg.de/en/forschung/computerscience/`.

In Sect. 2, we discuss the general LGCM and its assumptions. In Sect. 3, we describe two fitting functions, the Least Squares and the Minus 2 Log Likelihood,

and their implementation in our engine. The issues of starting values and non-admissible estimation areas is discussed. Section 4 describes the general simulation procedure. Data are generated in accordance with the LGCM and a set of known parameter values (we shall call "population values"), then selected following specific considerations about time sampling, and finally analyzed. Comparisons of slightly different LGCMs are presented to allow for inferential conclusions about single parameters of interest. Section 5 presents an illustration of the engine to investigate a particular set of parameters within the LGCM. A more detailed analysis of LGCM parameters under a wide variety of empirically plausible conditions is presented in [14, 16]. The present simulation serves illustrative purposes. Finally, in Sect. 6 we discuss our conclusions.

2 The Latent Growth Curve Model

Consider N units (e.g., persons) with K data points, corresponding to V variables measured at T time points (i.e., $K = VT$). The data points are obtained by applying a continuous function f_v defining the relations among P parameters, R Gaussian distributed random numbers, and T time points to each variable $v = 0, \ldots, V-1$. We call such a continuous function a *model*. Let $\mathcal{C}$ denote this function space. Then $\Sigma \in \mathcal{C}^{K \times K}$ denotes the covariance matrix of the data points with respect to the parameters. Likewise, $\mu \in \mathcal{C}^K$ denotes the vector of means with respect to the parameters. We denote the vector of parameters by $\overline{p}$.

We define here a particular linear model with equal interval measurement, where for each variable v and for each unit $i = 0, \ldots, N-1$ we consider a level $l_{v,i}$ and a slope $s_{v,i}$. A data point for a unit i is defined by

$$f_{t,v,i} = l_{v,i} + \frac{t}{T} \cdot s_{v,i} + err_{t,v,i} \tag{1}$$

where t is a time point (i.e., $t = 0, \ldots, T-1$) and $err_{t,v,i}$ is a normally distributed error term. While $err_{t,v,i}$ contains a time subscript t, indicating that its value changes across a unit's time series, both $l_{v,i}$ and $s_{v,i}$ are time invariant. All three terms are dependent on the unit i and variable v.

The means, variances, and covariances of $l_{v,i}$, $s_{v,i}$, and $err_{t,v,i}$ represent the parameters of the linear LGCM for each variable v (e.g., [26]). Assuming that the error components have mean zero, do not covary with any other parameter, and have a time-invariant variance, the parameters of the model are $2V$ means for all levels and slopes, $3V$ variances for all levels, slopes, and errors, and $\frac{2V(2V-1)}{2}$ covariances among all levels and slopes. The total number of parameters is $3V + \frac{2V(2V+1)}{2} = 2v^2 + 4v$.

μ and Σ are created with the matrix $\Lambda \in \mathbb{Q}^{K \times 2V}$, defined as

$$\Lambda_{ij} = \begin{cases} 1 & \lfloor i/T \rfloor = j \\ \frac{i-(j-V)T}{t} & \lfloor i/t \rfloor = j + V \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Λ will hence look like

$$\Lambda = \begin{pmatrix} 1 & 0 & & 0 & \frac{0}{T} & 0 & & 0 \\ 1 & 0 & & 0 & \frac{1}{T} & 0 & & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & & 0 & \frac{T-1}{T} & 0 & & 0 \\ 0 & 1 & & 0 & 0 & \frac{0}{T} & & 0 \\ 0 & 1 & & 0 & 0 & \frac{1}{T} & & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 1 & & 0 & 0 & \frac{T-1}{T} & & 0 \\ & & & & \vdots & & & \\ 0 & & & 1 & 0 & & & \frac{0}{T} \\ 0 & & & 1 & 0 & & & \frac{1}{T} \\ \vdots & \dots & \dots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & & & 1 & 0 & & & \frac{T-1}{T} \end{pmatrix} \tag{3}$$

To obtain the vector μ, we multiply Λ by the vector of the means of all levels and slopes:

$$\mu = \Lambda \left(\mu_{l_1}, \dots, \mu_{l_{V-1}}, \mu_{s_1}, \dots, \mu_{s_{V-1}}\right)^T \tag{4}$$

Let M denote the covariance matrix of levels and slopes in the same order. Σ is then obtainable by

$$\Sigma = \Lambda M \Lambda^T \tag{5}$$

Note that in the linear LGCM, all entries in μ and Σ are linear. Furthermore, the parameters in μ are only means of levels and slopes, while the parameters in Σ are covariances of levels, slopes, and error.

In sum, then, the usual application of LGCM in psychological research consists in analyzing N time series, one for each unit i of analysis, spanning over T time points, for a total of V variables. Researchers then wish to obtain information about the level $l_{v,i}$, the slope $s_{v,i}$, and the error $err_{t,v,i}$ for each variable v. The overall level means, slope means, level variances, slope variances, covariances among all levels and slopes, and error variances are the elements of $\overline{p}$.

3 Least Squares and Minus Two Log Likelihood Fitting Functions

To obtain the optimal parameter values by applying the LGCM to an $N \cdot K$ data matrix, *indices* are defined that mathematically define the distance between the observed data points and the expectations of the LGCM contingent upon the parameter values. Indices are norm functions on the parameters and the data, which are minimal iff the most suitable parameter values are estimated. We consider two indices, the *Euclidean distance* and the *Deviance*. We then discuss two associated fitting functions, which minimize these distances given the observed data and estimated parameters in $\overline{p}$. For the Deviance, we use an iterative procedure that needs starting values. Inadmissible estimation areas of the fitting functions are defined and finally we discuss how our engine handles them.

The Euclidean distance fit index defines the distance between the covariance matrix and mean vector of the data and the covariance matrix $\Sigma(\overline{p})$ and mean vector $\mu(\overline{p})$ predicted by the model given $\overline{p}$. Let S be the covariance matrix of the observed data and m be the mean vector of the observed data. Then, the Euclidean distance ls is defined as

$$ls = \sqrt{\sum_{i=0}^{K-1}(\mu(\overline{p})_i - m_i)^2 + \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\left(\Sigma(\overline{p})_{i,j} - S_{i,j}\right)^2} \tag{6}$$

So, the Euclidean distance is the Frobenius norm on the difference of covariance matrices plus the absolute value of the difference of the mean vectors. If ls is minimal, the Euclidean distance (in the $K^2 + K$ dimensional space) between (Σ, μ) and (S, m) is minimal. We call the point of global minimum the least square estimate.

The square root can be omitted for computing the least square estimate, and if the parameters are distinguished between those associated to the means (i.e., parameters appearing in μ, but not in Σ) and those associated to the variances–covariances (i.e., parameters appearing in Σ, but not in μ), both can be estimated separately. In linear models, the ls index is a polynomial of degree two, and hence its extremes are uniquely determined. ls can be obtained by computing the first two derivatives with respect to $\overline{p}$. The first derivative with respect to one parameter θ is

$$\frac{\partial ls^2}{\partial \theta} = \sum_{i=0}^{K} 2\left(\frac{\partial \mu_i}{\partial \theta} + m_i\right) + \sum_{i=0}^{K}\sum_{j=0}^{K} 2\left(\frac{\partial \Sigma_{i,j}}{\partial \theta} + S_{i,j}\right) \tag{7}$$

The second derivative with respect to θ_1 and θ_2 is

$$\frac{\partial^2 ls^2}{\partial \theta_1 \partial \theta_2} = \sum_{i=0}^{K} 2\frac{\partial^2 \mu_i}{\partial \theta_1 \partial \theta_2} + \sum_{i=0}^{K}\sum_{j=0}^{K} 2\frac{\partial^2 \Sigma_{i,j}}{\partial \theta_1 \partial \theta_2} \tag{8}$$

If the model is linear, $\frac{\partial^2 ls^2}{\partial\theta_1\partial\theta_2}$ is zero for $\theta_1 \neq \theta_2$ and constant otherwise, so a single step in Newton's Method (with any starting value, take $\overline{0}$ for simplicity) obtains the least square estimates.

The second fit index we consider is the *Deviance*, also commonly called the *Minus Two Log Likelihood* or $-2LL$. Lange [19] defined the *Deviance* to easily accommodate incomplete data patterns (for instance in pedigree analysis for behavioral genetics research). The *Deviance* is defined by

$$F(\overline{p}) = N \cdot K \cdot \ln 2\pi + \sum_{i=1}^{N} \ln|\Sigma(\overline{p})| + (x^{(i)} - \mu(\overline{p}))^T \Sigma(\overline{p})^{-1}(x^{(i)} - \mu(\overline{p})) \quad (9)$$

where $x^{(i)}$ is the data vector of the ith person. Since $\Sigma(\overline{p})$ is a covariance matrix, it follows that it must be positive definite and consequently has a positive determinant; $F(\overline{p})$ is considered undefined otherwise. Hence, the image of F is in $\mathbb{R}$.

The above definition of the *Minus Two Log Likelihood* easily allows handling of incomplete data by deleting, or filtering, the rows and columns in Σ and μ corresponding to missing data points (cf. [1, 27]). Formally, let $\mathcal{M}_i \subseteq \{1, \ldots, K\}$ denote the incomplete, or missing, values in $x^{(i)}$. Let $\Sigma_{\mathcal{M}_i}$ denote the matrix Σ with all columns and rows in $\mathcal{M}_i$ deleted, and $\mu_{\mathcal{M}_i}$ denote the vector of means with all elements in $\mathcal{M}_i$ deleted. Then,

$$F(\overline{p}) = N \cdot K \cdot \ln 2\pi + \sum_{i=1}^{N} \ln|\Sigma_{\mathcal{M}_i}(\overline{p})| + (x^{(i)} - \mu_{\mathcal{M}_i}(\overline{p}))^T \Sigma_{\mathcal{M}_i}(\overline{p})^{-1}(x^{(i)} - \mu_{\mathcal{M}_i}(\overline{p})) \quad (10)$$

Analogous to the ls fit index, we wish to obtain the minimal value of $F(\overline{p})$. We will call the $\overline{p}$ parameter values globally minimizing $F(\overline{p})$ the $-2LL$ estimates.

3.1 Minimization of the Fitting Functions

A convenient minimization method for the $-2LL$ index is Newton's Method applied to the first derivatives of $F(\overline{p})$, finding an extremum of $F(\overline{p})$ by searching for a common zero of the first derivatives. The following propositions are used:

$$\begin{aligned} \frac{\partial \ln|\Sigma|}{\partial\theta} &= \operatorname{trace}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta}\right) \\ \frac{\partial \operatorname{trace}(\Sigma)}{\partial\theta} &= \operatorname{trace}\left(\frac{\partial\Sigma}{\partial\theta}\right) \\ \frac{\partial\Sigma^{-1}}{\partial\theta} &= -\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta}\Sigma^{-1} \end{aligned} \quad (11)$$

We then obtain (see also [19])

$$\frac{\partial F}{\partial \theta} = \sum_{i=1}^{N} \text{trace}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta}\right) - 2\left(\frac{\partial \mu}{\partial \theta}\right)^{T}\Sigma^{-1}(x^{(i)}-\mu) \\ -(x^{(i)}-\mu)^{T}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta}\Sigma^{-1}(x^{(i)}-\mu) \tag{12}$$

The second derivatives can be computed in full generality by

$$\begin{aligned}\frac{\partial^2 F}{\partial \theta_1 \partial \theta_2} = \sum_{i=1}^{N} \text{trace}&\left(-\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_2}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_1} + \Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \theta_1 \partial \theta_2}\right) \\ &-2\left(\frac{\partial^2 \mu}{\partial \theta_1 \partial \theta_2}\right)^{T}\Sigma^{-1}(x^{(i)}-\mu) + 2\left(\frac{\partial \mu}{\partial \theta_1}\right)^{T}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_2}\Sigma^{-1}(x^{(i)}-\mu) \\ &+2\left(\frac{\partial \mu}{\partial \theta_1}\right)^{T}\Sigma^{-1}\left(\frac{\partial \mu}{\partial \theta_2}\right) + 2\left(\frac{\partial \mu}{\partial \theta_2}\right)^{T}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_1}\Sigma^{-1}(x^{(i)}-\mu) \\ &+(x^{(i)}-\mu)^{T}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_2}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_1}\Sigma^{-1}(x^{(i)}-\mu) \\ &-(x^{(i)}-\mu)^{T}\Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \theta_1 \partial \theta_2}\Sigma^{-1}(x^{(i)}-\mu) \\ &+(x^{(i)}-\mu)^{T}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_1}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_2}\Sigma^{-1}(x^{(i)}-\mu)\end{aligned} \tag{13}$$

Given that the above formulae are too complex for quick computation, we make use of the following simplifications:

$$\begin{aligned}\sum_{i=1}^{N} x^{(i)^T}\Sigma x^{(i)} &= \sum_{i=1}^{N}\sum_{j,k=1}^{K} x_j^{(i)}x_k^{(i)}\sigma_{j,k} \\ &= \sum_{j,k=1}^{K}\sigma_{j,k}\left(\sum_{i=1}^{N} x_j^{(i)}x_k^{(i)}\right)\end{aligned} \tag{14}$$

and similarly

$$\begin{aligned}\sum_{i=1}^{N} x^{(i)^T}\Sigma\,\mu &= \sum_{i=1}^{N}\sum_{j,k=1}^{K} x_j^{(i)}\mu_k\sigma_{j,k} \\ &= \sum_{j,k=1}^{K}\sigma_{j,k}\,\mu_k\sum_{i=1}^{N} x_j^{(i)}\end{aligned} \tag{15}$$

By computing the vector of all sums $X = \sum_{i=1}^{N} x^{(i)}$ and all moments $M_{jk} = \sum_{i=1}^{N} x_j^{(i)}X_k$ in advance, we get rid of the outer sums in all formulae.

In some models such as the LGCM, there are parameters associated to the means and others associated to the variances–covariances. That is, the set of parameters

$\theta_1, \ldots, \theta_P$ decomposes into two disjunct subsets $\theta_1, \ldots, \theta_l$ and $\theta_{l+1}, \ldots, \theta_P$, such that $\mu \in \mathbb{R}[\theta_1, \ldots, \theta_l]^K$ and $\Sigma \in \mathbb{R}[\theta_{l+1}, ..., \theta_P]^{K \times K}$. Thus, $\frac{\partial \Sigma}{\partial \theta_i}$ is zero if $i \leq l$ and $\frac{\partial \mu}{\partial \theta_i}$ is zero if $i > l$.

If we consider linear models, all entries in Σ and μ are linear, and therefore $\frac{\partial \Sigma}{\partial \theta_i \partial \theta_j} = \frac{\partial \mu}{\partial \theta_i \partial \theta_j} = 0$, regardless of i and j.

Consequently, the terms above simplify to the following five terms:
If $i \leq l$, that is, θ_i is a parameter associated to the means, then

$$\frac{\partial F}{\partial \theta_i} = 2 \left(\frac{\partial \mu}{\partial \theta_i} \right)^T \Sigma (X - K \cdot \mu) \tag{16}$$

for $i > l$

$$\begin{aligned} A &:= \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta} \Sigma^{-1} \\ \frac{\partial F}{\partial \theta_i} &= n \cdot \text{trace} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right) - M \odot A + 2 \mu^T A X - n \mu^T A \mu \end{aligned} \tag{17}$$

where M denotes the moment matrix above and $\odot$ the sum over component-wise products.
For $i \leq l$ and $j \leq l$, the second derivatives are

$$\frac{\partial^2 F}{\partial \theta_i \partial \theta_j} = 2K \left(\frac{\partial \mu}{\partial \theta_i} \right)^T \Sigma^{-1} \left(\frac{\partial \mu}{\partial \theta_j} \right) \tag{18}$$

for $i \leq l$ and $j > l$

$$\frac{\partial^2 F}{\partial \theta_i \partial \theta_j} = 2 \left(\frac{\partial \mu}{\partial \theta_i} \right)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1} (X - K \mu) \tag{19}$$

and finally for $i > l$ and $j > l$

$$\begin{aligned} A &:= \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \Sigma^{-1} \\ B &:= A \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} + \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} A \\ \frac{\partial^2 F}{\partial \theta_i \partial \theta_j} &= K \cdot \text{trace} \left(A \frac{\partial \Sigma}{\partial \theta_i} \right) + M \odot B - 2 \mu^T B X - n \mu^T B \mu \end{aligned} \tag{20}$$

These computations allow us to find the extremum of $F(\overline{p})$ by applying Newton's Method. In a simulation work, as is ours, data are generated by population values given a priori. Thus, it would be possible to take the population values as starting

values for the iteration process. In empirical research situations, however, the population values are not known, and an alternative way of providing starting values must be applied.

One common method is to use the least square estimates as starting values. In this line, the simulation engine first computes the least square estimates and uses them as starting values for the iterative process.

3.2 Inadmissable Estimation Areas

The $-2LL$ fit index is not defined on those parameter vectors p for which $\Sigma_{\overline{p}}$ is not positive definite. We call these areas *inadmissible estimation areas*. A strategy to avoid the estimation algorithm from falling into inadmissible estimation areas is to apply a *penalty function* (cf. [32]), which artificially increases the fit index adopted when the algorithm is approaching inadmissible areas. This will force the estimation algorithm away from inadmissible estimation areas.

Let Σ_k denote the upper $k \times k$ submatrix of Σ. We define the following penalty function *pen* which is added to the $F(\overline{p})$:

$$pen(\Sigma) = \sum_{k=1}^{K} p(\Sigma_k) \quad , \quad p(\Sigma_k) = \begin{cases} 0 & |\Sigma_k| \geq 0 \\ e^{-c|\Sigma_k|^2} - 1 & |\Sigma_k| \leq 0 \end{cases} \tag{21}$$

Thus,

$$\frac{\partial pen}{\partial \theta}(\Sigma) = \sum_{k=1}^{N} \frac{\partial p}{\partial \theta}(\Sigma_k) \tag{22}$$

and

$$\frac{\partial p}{\partial \theta}(\Sigma_k) = \begin{cases} 0 & |\Sigma_k| \geq 0 \\ -2c \frac{\partial |\Sigma_k|}{\partial \theta} |\Sigma_k| e^{-c|\Sigma_k|^2} & |\Sigma_k| \leq 0 \end{cases} \tag{23}$$

where

$$\frac{\partial |\Sigma_k|}{\partial \theta} = |\Sigma_k| \text{trace} \left(\Sigma_k^{-1} \frac{\partial \Sigma_k}{\partial \theta} \right) \tag{24}$$

We can hence simplify

$$-2c \frac{\partial |\Sigma_k|}{\partial \theta} |\Sigma_k| e^{-c|\Sigma_k|^2} = -2c|\Sigma_k|^2 \text{trace} \left(\Sigma_k^{-1} \frac{\partial \Sigma_k}{\partial \theta} \right) e^{-c|\Sigma_k|^2} \tag{25}$$

The derivative of *pen* is continuous. The second derivative for non-positive $|\Sigma_k|$ is

$$\begin{aligned}\frac{\partial^2 p}{\partial\theta_1\partial\theta_2}(\Sigma_k) &= |\Sigma_k|\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_2}\right)(-4c)|\Sigma_k|\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_1}\right)e^{-c|\Sigma_k|^2}\\ &\quad+(-2c)|\Sigma_k|^2\text{trace}\left(-\Sigma_k\frac{\partial\Sigma_k}{\partial\theta_2}\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_1}+\Sigma_k^{-1}\frac{\partial^2\Sigma_k}{\partial\theta_1\theta_2}\right)e^{-c|\Sigma_k|^2}\\ &\quad+(-2c)|\Sigma_k|^2\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_1}\right)(-2c)|\Sigma_k|^2\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_2}\right)e^{-c|\Sigma_k|^2}\\ &= (-2c)|\Sigma_k|^2e^{-c|\Sigma_k|^2}\left(\left(2-2c|\Sigma_k|^2\right)\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_2}\right)\right. \qquad (26)\\ &\quad\left.\text{trace}\left(\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_1}\right)+\text{trace}\left(-\Sigma_k\frac{\partial\Sigma_k}{\partial\theta_2}\Sigma_k^{-1}\frac{\partial\Sigma_k}{\partial\theta_1}+\Sigma_k^{-1}\frac{\partial^2\Sigma_k}{\partial\theta_1\theta_2}\right)\right)\end{aligned}$$

For non-negative $|\Sigma_k|$, the derivative is constant zero. For $|\Sigma_k| = 0$, the derivatives are zero by both definitions, so *pen* is twice differentiable. In the linear model, the first derivative of *pen* is zero for the mean parameters, because *pen* only depends on parameters associated to variances and covariances. Moreover, in the linear model, the second derivative of Sigma is zero for all parameters.

Computationally, the determinants of all Σ_k are computed first. If all of them are above zero, the penalty function and its derivatives are zero. The (relatively complex) computations for the derivatives of the penalty function are only applied otherwise.

However, when the matrix is not positive definite and relatively far away from the boundary of positive definiteness, the penalty function results in high values, which are likely to produce computational difficulties. We therefore do not want to rely solely on a penalty function. Moreover, Newton's Method on the first derivatives of $F(\overline{p})$ only reveals an extreme point, but not necessarily a minimum. To circumvent both problems, we introduce a modification of Newton's Method by extension of the idea of damping (cf. [10]). This modification has so far not been used elsewhere for parameter estimation. We choose a damping factor dependent on the two-dimensional situation along the gradient. We quickly summarize the idea.

In the original Newton's Method, to find a common zero of $f_1, \ldots, f_K : \mathbb{R}^K \to \mathbb{R}$, in each step the *Jacobian* J of $f_1, \ldots, f_K$ defined by

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial p_1} & \cdots & \frac{\partial f_K}{\partial p_1} \\ \cdots & \ddots & \cdots \\ \frac{\partial f_1}{\partial p_K} & \cdots & \frac{\partial f_K}{\partial p_K} \end{pmatrix} \qquad (27)$$

is computed. If $p^{(i)}$ denotes the position of the ith step, then

$$p^{(i+1)} = p^{(i)} - J(p^{(i)})^{-1} f(p^{(i)}) \qquad (28)$$

To i mprove the quality of $p^{(i+1)}$, replace this line by

$$p^{(i+1)} = p^{(i)} + \lambda J(p^{(i)})^{-1} f(p^{(i)})$$

where $\lambda \in \mathbb{R}$ is a parameter that can be freely chosen. The original Newton's Method is achieved by setting $\lambda = -1$. We make a successive search for the best λ by the following method:

Let $p_\lambda := p^{(i)} + \lambda J(p^{(i)})^{-1} f(p^{(i)})$ and $f_\lambda = f(p_\lambda)$. Consider three values $\lambda_0 < \lambda_1 < \lambda_2$ initially set to -1, 0, 1. By decreasing λ_0, respectively increasing λ_2, we change the three values until f_{λ_1} is the minimum of the three values. Because the direction of the gradient $J(p^{(i)})^{-1} f(p^{(i)})$ points toward an extremum, we expect to find three values with the requested condition quickly. Otherwise, we take the preceding best value of λ corresponding to the lowest f_λ.

When $\lambda_0 < \lambda_1 < \lambda_2$ with $f_{\lambda_1} < f_{\lambda_0}$ and $f_{\lambda_1} < f_{\lambda_2}$ are found, we check whether $f_{\lambda_0} > f_{\lambda_2}$. Let λ_i correspond to the higher of the two, and let $\lambda_4 := \dfrac{\lambda_i + \lambda_1}{2}$ be the mean of λ_i and λ_1. We then check whether f_{λ_4} or f_{λ_1} is smaller and repeat the process with the corresponding λ and its two neighboring λ. These three λs again respect the condition that the middle λ corresponds to the lowest value. We repeat this process until λ_0, λ_1, and λ_3 only differ by a small a priori distance and continue with Newton's Method on p_{λ_1}.

4 General Simulation Procedure

To test certain statistical properties of the LGCM, we will create data points with an LGCM and a given set of known parameters (similarly to [14, 16]). Here we explain how the completed data points were created, how the data points were selected to match substantive questions of interest, and the evaluation criteria we computed to appraise the quality of the statistical methods (in particular how the parameter estimates themselves can be compared to the initial population parameter values).

4.1 Data Generation

Consistent with the LGCM, we specify a priori a variance–covariance matrix $M \in \mathbb{R}^{2V \times 2V}$ of the levels and slopes, vector $\mu \in \mathbb{R}^{2V}$ of the means of the levels and slopes, and an error variance θ uncorrelated with any other variable. From these initial parameters, we generate general level $l_{v,i}$ and slope $s_{v,i}$ scores, for variable v and unit of analysis i, such that the first and second moments of $l_{v,i}$ and $s_{v,i}$ correspond to μ and M, respectively.

To this end, we perform a Cholesky decomposition of the covariance matrix M of the parameters, i.e., we find a lower left triangular matrix C, such that $CC^T = M$. Since M is a covariance matrix, it is positive definite, and thus the Cholesky Decomposition exists. The matrix C can be computed recursively by

$$c_{i,j} = \begin{cases} \dfrac{m_{i,j} - \sum_{k=1}^{i-1} c_{k,i} c_{k,j}}{c_{i,i}} & i < j \\ \sqrt{m_{i,j} - \sum_{k=1}^{i-1} c_{k,i} c_{k,j}} & i = j \\ 0 & i > j \end{cases} \tag{29}$$

Let $r \in \mathbb{R}^N$ be N Gaussian-distributed random variables. The levels and slopes can then be obtained by taking

$$(l_{0,0}..., l_{V-1,N-1}, s_{0,0}..., s_{V-1,N-1}) = Cr + \mu$$

$l_{v,i}$ and $s_{v,i}$ are then normally distributed with means μ and variance–covariance matrix M, as can be checked

$$\begin{aligned} \int_r Cr(Cr)^T \omega(r) &= \int_r Crr^T C^T \omega(r) \\ &= C \left(\int_r rr^T \omega(r) \right) C^T \\ &= CIC^T \\ &= M \end{aligned} \tag{30}$$

where

$$\int_r f(r)\omega(r) := \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (2\pi)^{\frac{-n}{2}} e^{\frac{-x_1^2}{2} - \dots \frac{-x_n^2}{2}} f(r) dr_1 \cdots dr_n$$

as a short notation for integrals over some Gaussian distribution.

The error term $err_{t,v,i}$ can be computed by multiplying a Gaussian-distributed variable by θ, independently for all t, v, and i.

Hence, in the end we generate N level scores $l_{v,i}$, slope scores $s_{v,i}$, and error scores $err_{t,v,i}$, which all correspond to the initial population LGCM parameters M, μ, and θ. These values are finally combined according to the LGCM equation (1) to obtain K final data points for each unit of observation.

4.2 Data Selection

Let $x^{(i)}$ be the data points created as described above given the population variance–covariance matrix Σ, the population mean vector μ, and the population error variance–covariance matrix θ. At this point we may apply the estimation procedures explained above to obtain the LGCM parameter values from the observed data $x^{(i)}$. Alternatively, if we are interested in specific statistical properties of the LGCM, we

may *select* some of the data from $x^{(i)}$. For instance, we may test the robustness of LGCM to incomplete data under certain conditions (e.g., [28, 29, 12]).

One dimension of interest in our simulation is the time interval spanned by the data points T. Large scale longitudinal studies, or panels, are very expensive and laborious, and consequently typically last less than a decade (although notable exceptions exist). Of interest to many applied researchers is the necessary duration of a longitudinal study in order to detect reliable variance in change (cf. [15]). In the context of LGCM, this question translates into the number of longitudinal data points necessary to detect reliable variance in the slope scores $s_{\nu,k}$. Other LGCM parameters defined in Σ, μ, or θ can of course be examined.

To select the data to be examined, we perform a selection operation on the complete data set $x^{(i)}$ of each unit of observation i. Let A be the full data set of all data points for each unit of observation. For each selection condition, we select (randomly or according to some predefined criterion, such as the number of longitudinal measurements) a subset $J \subseteq A$ of the observed data points. We then restrict the data vector $x^{(i)}$ of each unit of observation, the matrix Σ, the vector μ, and the matrix θ to the rows and columns corresponding to the indices in J (for each variable ν).

This selection operation allows testing the robustness of LGCM to incomplete data. The LGCM is then applied to the resulting subset of data J and the overall fit of the model (ls or $Deviance$) and the resulting parameter estimates $\overline{p}$ are evaluated in light of prespecified criteria.

4.3 Evaluation Criteria

We specified several evaluation criteria based on the statistical distribution of the χ^2 statistic and its degrees of freedom. For each LGCM computed on every generated data set, we obtain the parameters and evaluation criteria summarized in Table 1.

Table 1 Parameters and evaluation criteria

1	Least square estimates
2	$-2LL$ estimates
1	$-2LL$ value for the estimates
2	$-2LL$ value for the saturated model
3	$-2LL$ value for the population parameters
4	$-2LL$ value for the independent model with free means
5	$-2LL$ value for the independent model with averaged means
6	$-2LL$ value for the parameters of the free model
7	SRMR for the variance–covariance matrix
8	SRMR for the mean vector

Here, we have (a) the following estimates:

1. The estimated parameters for the minimal Least Square index. This includes all elements of Σ and μ estimated with the ls procedure specified above.

2. The estimated parameters for the minimal $-2LL$ index. This includes all elements of Σ and μ estimated with the $-2LL$ procedure specified above.

and (b) the following fit indices:

1. The $-2LL$ value of the $-2LL$ estimates. This is the actual value of the $-2LL$ fit index corresponding to the parameter estimates obtained with the $-2LL$ procedure (cf. the estimates in point 2 above). In this application, the model is a LGCM with all parameters freely estimated.
2. The $-2LL$ value for the saturated model. This is the $-2LL$ value for the model estimating one parameter for each unknown, that is, one parameter for each element in Σ and in μ. This model is the least parsimonious and yields the best fit to the data. Indeed, to obtain this $-2LL$ we substituted Σ with S and μ with m in the equation for $-2LL$.
3. The $-2LL$ value for population parameters. This corresponds to the $-2LL$ fit index when the parameters are not estimated, but fixed at the known population values, with which the data were generated in the first place. In this application, the model is a LGCM with all parameters fixed (hence not estimated) to the initial population values.
4. The $-2LL$ value for the independent model with free means. This is a common baseline comparison model in the structural equation modeling literature (e.g., [7]). This $-2LL$ value is the value obtained when the independence model, rather than the LGCM, is compared to the observed data. In the independence model, all longitudinal measures are posited independent of each other, but with possibly different variance values. The model expectation variance–covariance matrix S is hence a diagonal matrix, where all covariances are equal to zero. To separate the effects due to the variance–covariance matrix Σ from those of the mean vector μ, this first independence model estimates all mean values separately, so that m counts $(V - 1) \cdot (T - 1)$ parameters.
5. The $-2LL$ value for the independent model with averaged means. This fit index is equivalent to the previous, except that it is also restrictive on the mean structure. Here only $(V - 1)$ parameters are estimated for the means, that is, only one mean value for each variable v. Hence this model posits independence among the longitudinal measurements and no average longitudinal change.
6. The $-2LL$ value of the free model. To reach statistical conclusions about a specific LGCM parameter (e.g., the correlation between two variables' slope scores), the initial LGCM model, whose $-2LL$ fit is (1), will be constrained with a value of 0 for that specific parameter. The statistical inference about that parameter can be based on the $-2LL$ contribution due to that parameter, which is the difference between the $-2LL$ in (1) and the $-2LL$ of the free model, in which the parameter is freely estimated rather than constrained at 0.
7. The Standardized Root Mean Residual (SRMR) for the variance–covariance matrix. This fit index is similar to the squared Euclidean distance between the standardized data variance–covariance matrix S and the model expected

variance–covariance matrix Σ with the parameter values $\overline{p}$. This SRMR is defined by

$$SRMR_{\text{cov}}(\overline{p}) = \sum_{i=0}^{K}\sum_{j=0}^{K}\left(\frac{\Sigma(\overline{p})_{ij}}{\sqrt{\Sigma(\overline{p})_{ii}\,\Sigma(\overline{p})_{jj}}} - \frac{S_{ij}}{\sqrt{S_{ii}\,S_{jj}}}\right)^2 \tag{31}$$

8. The Standardized Root Mean Residual (SRMR) for the mean vector. This SRMR ignores all variance–covariance information and assesses the squared Euclidean distance between the mean vectors μ and m:

$$SRMR_{\text{mean}}(\overline{p}) = \sum_{i=0}^{K}(\mu(\overline{p})_i - m_i)^2 \tag{32}$$

All models described above are statistically nested in the free model, meaning that the parameters estimated by each model are a subset of the parameters of the free model. The free model will always obtain a Deviance, which is smaller or equal to the any other model, because it describes the overall data structure better or as equally well as any other model. The $-2LL$ difference between any other model and the free model is distributed as a χ^2 statistic with as many degrees of freedom as the difference between the number of parameters estimated by the two models, because they are statistically nested. Differences of $-2LL$ statistics can be re-expressed as Comparative Root Mean Square Error of Approximation (cf. [8]), defined as

$$CRMSEA = \sqrt{\frac{\max[(\frac{\Delta\chi^2 - \Delta df}{N-1}), 0]}{\Delta df}} \tag{33}$$

where $\Delta\chi^2$ corresponds to the difference in χ^2 values and Δdf to the difference in degrees of freedom (df) between the two nested models.

4.4 Summarizing the Simulation Procedure

The total simulation procedure is illustrated in Fig. 1. First, the population values and the model for data creation are used to generate the data set as described in Sect. 4.1. Possibly, specific data points are selected as described in Sect. 4.2, and the model is restricted by constraining specific parameters of focus. The restricted model is then applied to the selected data set to minimize the least squares or the $-2LL$ index and to obtain the estimated parameters. The estimated parameters may then be compared to the original parameters (i.e., the population values) or by certain evaluation criteria to finally ascertain their quality, as described in Sect. 4.3.

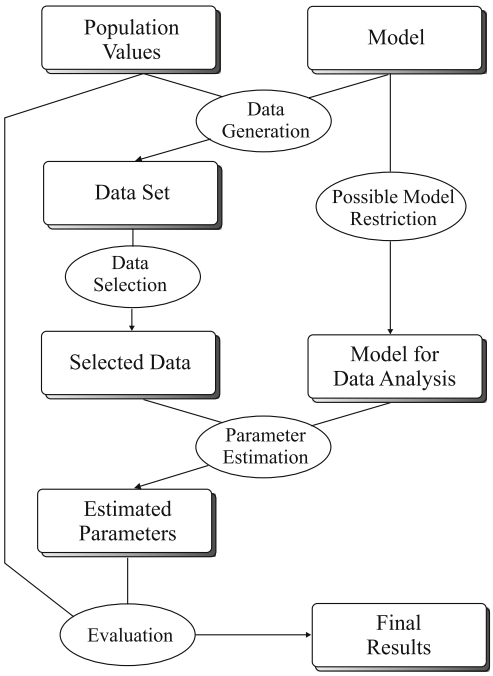


Fig. 1 Representation of simulation strategy

5 An Illustration

In this section, we present an illustration of our engine for data generation plus least squares and full information maximum likelihood estimation to test statistical properties of LGCMs. We will limit the analyses to two LGCM parameters. A more extensive analysis of LGCM statistical properties is provided in [14, 16].

In this application, we were particularly interested in testing the power of LGCMs in estimating variances and covariances of the slope components. Much recent work in our main research field, cognitive aging, has focused on interindividual differences in change, or differential change, and relationships of change over time. General salient questions in adult cognitive development concern whether aging individuals change similarly or display subgroups according to their developmental patterns (e.g., the successful aging paradigm proposed by [35, 3]) and whether changes in one domain, such as cognitive abilities, are related to changes in other domains, such as sensory functions (e.g., [9]).

5.1 Population Parameters

Based on existing research examples (cf. [14, 16]), we generated data on $V = 2$ variables for $N = 200$ and $N = 500$ units of analysis over $T = 20$ time points. We

chose the following variance–covariance matrix M and mean vector μ for levels and slopes (in the order level of variable one, level of variable two, slope of variable one, and slope of variable two):

$$M = \begin{pmatrix} 100 & & & \\ 50 & 100 & & \\ X & 0 & \mathbf{50} & \\ 0 & X & \mathbf{25} & \mathbf{50} \end{pmatrix} \mu = \begin{pmatrix} 50 \\ 50 \\ -20 \\ -20 \end{pmatrix} \tag{34}$$

In this example, we were interested in the effect of the within-variable level-slope covariance X on the power to detect the covariance between the two slopes. To this end, we varied X from -21.21 to 0.00 to 21.21, which corresponds to varying the within-variable level-slope correlation from -0.3 to 0.0 and 0.3, respectively. The error variance of both variables was set to 10, which is equivalent to an initial reliability of 0.91 for both variables. This corresponds to the reliability of good cognitive tests (e.g., [23]).

5.2 Data Selection

Two hundred replicate data sets per combination of population parameters are generated and successively altered and analyzed. The data selection concerned the time dimension, or length of the individual time series, only. The full data sets generated in the first step simulate cognitive aging studies lasting $T = 20$ epochs (e.g., weeks, months, or, relevant to the choice of the population parameters, years). In empirical terms, it is very laborious and problematic for several reasons to repeatedly measure a sample for 20 years. One practical question for researchers is then: For how many years, or on how many occasions, do I need to measure my participants to detect reliable variance in change and reliable covariance in change with LGCM analyses (as a function of sample size, real variance and covariance in change, and variables' reliabilities)?

Hence we selected data according to the time dimension, by retaining 3 ($t = 0, 2, 4$), 4 ($t = 0, 2, 4, 6$), 5 ($t = 0, 2, 4, 6, 8$), 6 ($t = 0, 2, 4, 6, 8, 10$), 10 ($t = 0, 2, 4, 6, 8, 10, 12, 14, 16, 18$), and all 20 (from 0 to 19, by increments of 1) repeated measures. This condition will allow comparing the power to detect reliable variance in change and covariance in change as a function of study length, sample size, and the effect size of within-variable level-slope correlations (X in (34)), accounting for variables' reliabilities and the effect sizes of the variances and means of levels and slopes.

5.3 Parameters of Focus

In this illustration, we focus on the two variances of slope scores $Variance(s_{1,k})$ and $Variance(s_{2,k})$ and the covariance between them, $Covariance(s_{1,k}; s_{2,k})$. The three parameters are boldfaced in the matrix representation of M in (34). During the parameter estimation procedure, we will hence solve three models: (M1) the LGCM with all parameters freely estimated; (M2) an LGCM with the covariance of slope scores fixed at 0, hence not estimated; and (M3) an LGCM with both variances of slope scores and their dependent covariances fixed at 0. Model (M1) estimates the total number of parameters of a bivariate LGCM, that is 16, model (M2) estimates 15 parameters, and model (M3) estimates 9 parameters.

Besides the evaluation of the three models by means of the fit indices described in 4.3, relative model comparisons are possible. Models (M2) and (M3) are statistically nested within (M1) and model (M3) is statistically nested in (M2), so that statistical pairwise model comparisons are justified.

5.4 Definition of Power

The main dependent variable of our illustration concerns the power of LGCMs to correctly reject the null hypothesis that the parameters of focus are equal to 0 (when their analogous population parameters are different from 0). In this illustration, power of the parameters of focus is defined for all combinations of population parameters, because $Variance(s_{1,k}) = Variance(s_{2,k}) = 50$ and $Covariance(s_{1,k}; s_{2,k}) = 25$ in all combinations of population parameters.

To define power in our simulation, we computed for all 200 replicates of each combination of population parameters two statistical comparisons: We compared models (M1) and (M2) to calculate the loss in $-2LL$ fit due to not estimating the $Covariance(s_{1,k}; s_{2,k})$ and models (M1) and (M3) for the loss in fit attributable to the $Variance(s_{1,k})$ and $Variance(s_{2,k})$ and all dependent covariances (which are not defined when their relative variances are zero). The two differences in fit are distributed as a χ^2 with 1 and 7 degree(s) of freedom, respectively. A significant model comparison is obtained when the comparison χ^2 value is greater, at an alpha level of 0.05, than 3.84 and 14.07, respectively. Power is then defined as the ratio of significant model comparisons out of the total 200 for each combination of population parameters.

5.5 Results

The power estimates for detecting the $Variance(s_{1,k})$ and $Variance(s_{2,k})$ are plotted in Fig. 2, and those for $Covariance(s_{1,k}; s_{2,k})$ in Fig. 3.

In general, the within-variable level-slope correlation seems to affect power to detect the variance of slopes only when three occasions are retained with $N = 200$

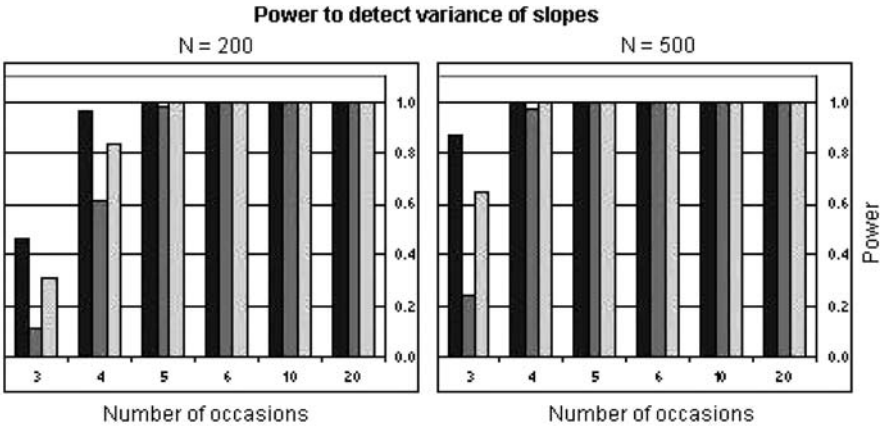


Fig. 2 Power to detect variances of slopes as a function of within-variable level-slope correlation (*left*, *dark bar* $r = -0.30$, *middle*, *grey bar* $r = 0$, *right*, *white bar* $r = 0.30$) and occasions retained

or $N = 500$ and with four occasions with $N = 200$. When five or more occasions are retained, power is very high with both sample sizes across the three values of the within-variable level-slope correlation.

Power to detect the covariance of slopes is acceptable with six occasions or more when $N = 200$ and five or more occasions with $N = 500$, independently of the within-variable level-slope correlation.

To test formally the effects observed with the barplots, we tested the differences between the analogous power columns of both sample sizes for significance. The event to successfully reject the null hypothesis is binary distributed with the power

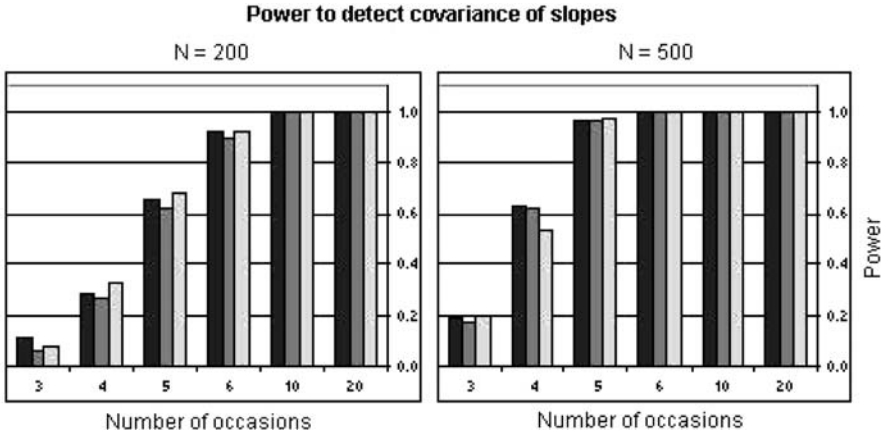


Fig. 3 Power to detect covariance of slopes as a function of within-variable level-slope correlation (*left*, *dark bar* $r = -0.30$, *middle*, *grey bar* $r = 0$, *right*, *white bar* $r = 0.30$) and occasions retained

as probability for rejection. Thus, the number of rejections in 200 replicates is binomially distributed with the power equal to the probability. Two cells are therefore significantly different to an α level if the probability of being drawn from the same binomial distribution is less than α.

This analysis confirmed our assumptions that occasions strongly affect the power to detect both variables' slope variance and the across-variable slope covariance. All comparisons between two analogous cells of the same parameter but differing with respect to occasions were significantly different ($p < 0.001$), unless the power was maximum (i.e., 100%). Likewise, the effect of sample size was equally significant ($p < 0.001$) for all analogous cells differing only with respect to sample size, unless power was 100%.

The within-variable level-slope covariance had no significant effects on the power to detect the across-variable covariance of slopes at an $\alpha = 0.01$ level. Yet, the within-variable level-slope covariance had a significant effect on the power to detect the two slope variances. For the power to detect variance in change, all cells different only in level-slope covariance were highly significantly different ($\alpha < 0.001$, unless power was 100%).

In sum, the simulation showed that detection power of variance is higher with a positive within-variable correlation between level and slope, and even higher when this correlation is negative. The power of across-variable covariance of change, on the other hand, does not appear to be significantly affected by the within-variable level-slope correlations.

6 Discussion and Outlook

In this chapter, we presented a simulation procedure for testing statistical properties of Latent Growth Curve Models (LGCM). In the application, we applied the procedure to study the power of LGCMs to detect variances and covariances of linear change. The simulation engine (`http://www.mpib-berlin.mpg.de/en/forschung/computerscience/`.) produced data according to a linear LGCM with different parameters, then selected those data sets, and finally analyzed them under different parameter restrictions to compute nested model comparisons focused around parameters of interest (variances in and covariance of change).

To estimate the LGCM parameters for each generated data set, we provided some technically equivalent transformations of the derivatives of the Minus Two Log Likelihood index, which allowed us quickly finding minimal points by a variant of Newton's Method. In this manner, we were able to avoid areas of the parameter space that are inadmissible for covariance matrices and to separate minima from maxima.

In the illustration, we showed that the power to detect variances of change in a LGCM is dependent on the within-variable level-slope covariance, while the power to detect across-variable covariance of changes in a LGCM apparently is not. A possible explanation of this effect can be found in detail in [16].

In short, there is more than one possible statistical test for the variance in change. One may nest a model with both variances in change constrained to be zero within an unrestricted model. This will lead to a 2-degree-of-freedom chi square comparison. Alternatively, one may also compare a model with both variances in change fixed to zero as well as all related covariances, which corresponds to a 7-degree-of-freedom chi square comparison (2 variances and 7 covariances). In [16], we showed that while the latter method is superior when the real covariances associated to the change factors in the population are zero, the former is superior when those covariances are different from zero. Because in the present illustration we applied the former method, the resulting power to detect variances in change increased with higher within-variable level-slope covariance.

In future work, we intend to elaborate our research of simulation methods and expand the simulation engine to cope with the incongruence due to creating data with one LGCM specification and subsequently analyzing those data with a different LGCM specification. Also, the effects of more complex data selection strategies will be addressed.

LGCMs have become a prominent method of data analysis in much psychological researchs. These models are appealing because they (a) allow disentangling level from change in information; (b) allow specifying a wide variety of pre-defined change patterns (e.g., polynomial, exponential, and Gompertz) or estimating the change pattern empirically from the data analyzed; (c) allow analyzing all data available, even in the presence of incomplete data, as long as the missing at random assumption is met; and most importantly (d) have contributed significantly to the advancement of theoretical knowledge about the cognitive aging literature.

The study of statistical behaviors of LGCMs is however still a very active research field. Although much cognitive aging literature focuses on change parameters, especially variance and covariance, the field as a whole still does not know the limits of LGCMs. We showed that even under ideal and empirically unrealistic assumptions about the data (e.g., group homogeneity with respect to the change phenomenon examined, nonexistent longitudinal dropout, and correct a priori specification of the change function) certain LGCM parameters of chief substantive importance are estimated with low to very low power.

Simulation studies such as this allow us furthering our knowledge about the limits and tenability of LGCMs under given research situations. We believe that much more research is needed to persuade LGCM users not to rest on substantive findings, which might be invalid because of inherent LGCM lack of power under specific conditions, most of which still in need of being discovered.

References

1. Arbuckle, J.L. Full information estimation in the presence of incomplete data. In: G.A. Marcoulides, R.E. Schumacker (Eds.), Advanced Structural Equation Modeling: Issues and Techniques (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (1996).

2. Arbuckle, J.L., Wothke, W. Amos 4.0. User's Guide. Chicago, IL: SmallWaters Corporation (1995).
3. Baltes, P.B., Baltes, M.M. Successful Aging: Perspectives from the Behavioral Sciences. Cambridge, UK: Cambridge University Press (1990).
4. Baltes, P.B., Nesselroade, J.R. History and rationale of longitudinal research. In J.R. Nesselroade, P.B. Baltes (Eds.), Longitudinal Research in the Study of Behavior and Development (pp. 1–39). New York: Academic Press, Inc. (1979).
5. Baltes, P.B., Reese, H.W., Nesselroade, J.R. Life-Span Developmental Psychology: Introduction to Research Methods. Monterrey, CA: Brooks/Cole (1977).
6. Bentler, P.M. EQS Program Manual. Encino, CA: Multivariate Software, Inc. (1995).
7. Bollen, K.A. Structural Equations with Latent Variables. New York: John Wiley (1989).
8. Browne, M., Cudeck, R. Alternative ways of assessing model fit. In K.A. Boolen, J.S. Long (Eds.), Testing Structural Equation Models (pp. 136–162). Newbury Park, CA: Sage Publications, Inc. (1993).
9. Deary, I.J. Sensory discrimination and intelligence: Postmortem or resurrection? American Journal of Psychology, 107:95–115 (1994).
10. Deuflhard, P., Hohmann, A. Numerische Mathematik [Numerical mathematics]. Berlin, Germany: Walter de Gruyter (1993).
11. Fan, X. Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. Structural Equation Modeling, 10:380–400 (2003).
12. Ghisletta, P. A simulation analysis of alternative methods to correct for selective dropout in longitudinal studies. Unpublished doctoral thesis. University of Virginia, Charlottesville, Virginia (1999).
13. Ghisletta, P., Lindenberger, U. Static and dynamic longitudinal structural analyses of cognitive changes in old age. Gerontology, 50:12–16 (2004).
14. Hertzog, C., Lindenberger, U., Ghisletta, P., Oertzen, T. On the power of multivariate latent growth curve models to detect correlated change. Psychological Methods, 11(3):244–252 (2006).
15. Hertzog, C., Nesselroade, J.R. Assessing psychological change in adulthood: An overview of methodological issues. Psychology and Aging, 18:639–657 (2003).
16. Hertzog, C., von Oertzen, T., Ghisletta, P., Lindenberger, U. Evaluating the power of latent growth curve models to detect individual differences in change. Structural Equation Modeling, 15:541–563 (2008).
17. Jöreskog, K.G., Sörbom, D. LISREL 8. User's Reference Guide. Chicago, IL: Scientific Software International (1996).
18. Laird, N.M., Ware, J.H. Random-effects models for longitudinal data. Biometrics, 38: 963–974 (1982).
19. Lange, K., Westlake, J., Spence, M.A. Extensions to pedigree analysis. iii. variance components by the scoring method. Annals of Human Genetics, 39:485–491 (1976).
20. Lindenberger, U. Lifespan theories of cognitive development. In N. Smelser, P. Baltes (Eds.), International Encyclopedia of the Social and Behavioral Sciences (pp. 8848–8854). Oxford: Elsevier Science (2001).
21. Lindenberger, U., Chicherio, C. Développement intellectuel au cours du cycle de vie : Sources de variabilité et niveaux d'analyse. L'Année Psychologique, 108:757–793 (2008).
22. Lindenberger, U., Ghisletta, P. Modeling longitudinal changes in old age: From covariance structures to dynamic systems. In R.A. Dixon, L. Bäckman, L.G. Nilsson (Eds.), New Frontiers in Cognitive Aging (pp. 199–216). Oxford, UK: Oxford University Press (2004).
23. Lindenberger, U., Mayr, U., Kliegl, R. Speed and intelligence in old age. Psychology and Aging, 8:207–220 (1993).
24. Lindenberger, U., Oertzen, T. Variability in cognitive aging: From taxonomy to theory. In F. Craik, E. Bialystok (Eds.), Lifespan Cognition: Mechanisms of Change (pp. 297–314). Oxford: Oxford University Press (2006).

25. Marsh, H., Hau, K.T. Multilevel modeling of longitudinal growth and change: Substantive effects or regression toward the mean artifacts? Multivariate Behavioral Research, 37:245–282 (2001).
26. McArdle, J.J. Latent growth within behavior genetic models. Behavior Genetics, 16:163–200 (1986).
27. McArdle, J.J. Structural factor analysis experiments with incomplete data. Multivariate Behavioral Research, 29:409–454 (1994).
28. McArdle, J.J., Hamagami, F. Modeling incomplete longitudinal data using latent growth structural equation models. In L.M. Collins, J.L. Horn (Eds.), Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions (pp. 276–304). Washington, DC: American Psychological Association (1991).
29. McArdle, J.J., Hamagami, F. Modeling incomplete cross-sectional and longitudinal data using latent growth structural models. Experimental Aging Research, 18:145–166 (1992).
30. McArdle, J.J., Hamagami, F. Multilevel models from a multiple group structural equation perspective. In G.A. Marcoulides, R.E. Schumaker (Eds.), Advanced Structural Equation Modeling. Issues and Techniques (pp. 89–124). Mahwah, NJ: Lawrence Erlbaum Associates (1996).
31. Muthén, L.K., Muthén, B.O. MPlus User's Guide. Los Angeles, CA: Muthén and Muthén (1998).
32. Neale, M.C., Boker, S.C., Xie, G., Maes, H.H. Mx: Statistical Modeling (5th edn.). Richmond: Medical College of Virginia (1999).
33. Raudenbush, S.W. Toward a coherent framework for comparing trajectories of individual change. In L.M. Collins, A.G. Sayer (Eds.), New Methods for the Analysis of Change (2nd edn., pp. 33–64). Washington, DC: American Psychological Association (2000).
34. Rovine, M., Molenaar, P.C.M. The covariance between level and shape in the latent growth curve model with estimated basis vector coefficients. Methods of Psychological Research Online, 3:95–107 (1998).
35. Rowe, J.W., Kahn, R.L. Human aging: Usual and successful. Science, 237:143–149 (1987).
36. Singer, T., Verhaeghen, P., Lindenberger, U., Baltes, P. The fate of cognition in very old age: Six-year longitudinal findings in the Berlin aging study (BASE). Psychology and Aging, 18:318–331 (2003).

Conflicting Constraints in Resource-Adaptive Language Comprehension

Andrea Weber, Matthew W. Crocker, and Pia Knoeferle

1 Introduction

The primary goal of psycholinguistic research is to understand the architectures and mechanisms that underlie human language comprehension and production. This entails an understanding of how linguistic knowledge is represented and organized in the brain and a theory of how that knowledge is accessed when we use language. Research has traditionally emphasized purely linguistic aspects of on-line comprehension, such as the influence of lexical, syntactic, semantic and discourse constraints, and their time-course. It has become increasingly clear, however, that non-linguistic information, such as the visual environment, are also actively exploited by situated language comprehenders. The wealth of informational resources which are potentially relevant to situated language comprehension raise a number of important questions. To what extent are the mechanisms underlying comprehension able to exploit linguistic and non-linguistic information on-line, and how do people adapt to the availability or non-availability of contextual information?

We begin below, with a brief summary of several important aspects of human language comprehension, including its incremental and even anticipatory nature, and its sensitivity to accrued linguistic experience. We then present a range of experimental findings which reveal the ability of comprehenders to rapidly adapt to diverse linguistic and non-linguistic constraints. To better understand this apparently seamless ability of the human language processing faculty to integrate diverse cues, including linguistic context, intonation, world knowledge, and visual context, many of the experiments are designed so as to better understand the relative priority of these constraints when they are pitted against each other. The findings conspire to paint a picture in which purely linguistic constraints, long thought to identify the core of sentence comprehension mechanisms, can in fact be overridden by highly contextual aspects of the situation, such as the intonation contour of a particular utterance, semantic expectations supported by the visual scene, and indeed events going on in the scene itself.

A. Weber (✉)
Max-Planck-Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands
e-mail: andrea.weber@mpi.nl

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_7,

1.1 Incrementality

A basic finding is that human sentence processing is incremental. That is, humans structure and interpret the words of an utterance as they are perceived rather than store them as a list to be combined later. A seminal finding for incrementality was published in 1973 by Marslen-Wilson [24], who showed in a speech-shadowing experiment that both syntactic and semantic information are available to participants as they repeat the speech they hear; constructive errors were usually grammatically suitable with respect to the preceding context, even for shadowers who repeated the speech input with a minimal time-lag. This suggests that the shadowers' performance was based on a syntactic analysis of the ongoing speech stream. Since that time, empirical support for the claim that language comprehension takes place incrementally is overwhelming. Evidence from eye-tracking has shown that people not only rapidly map the unfolding words onto visually present objects, but that they also structure the words of an utterance into a connected interpretation as they are encountered (e.g., [35]), and they even have expectations about the words they predict to come (e.g., [2]).

However, while incremental processing and interpretation ensures real-time understanding, it brings with it additional challenges. Sequences are often ambiguous; that is, they are compatible with more than one well-formed structural representation. For example, in the sentence beginning *Betty knew Monica's date ...*, *Monica's date* could be the direct object of *knew* or could become the subject of a clausal complement (*Betty knew Monica's date had bought flowers*). Disambiguating information may occur in later parts of the sentence, but due to incrementality, processing must proceed before such relevant information becomes available. A great deal of research has therefore focused on the processing of local ambiguities as a means for investigating the kinds of information and strategies listeners employ during the earliest stages of sentence processing [13, 7, 27].

1.2 Multiple Constraints

An abundance of empirical studies have specified the different information sources that are used on-line for ambiguity resolution in sentence processing. For example, it has been shown numerous times that the human parser resolves structural ambiguities using a set of processing preferences. One of these is a preference to always build the simplest structure; a direct object attachment of a postverbal noun phrase (NP) is, for example, less complex than an attachment that would require the additional structure associated with a complement clause (the parser thus prefers to analyze *Monica's date* as direct object in *Betty knew Monica's date*). This preference is known as the Minimal Attachment principle [13].

Besides purely structural information, prior linguistic experience has been shown to play an important role in human sentence processing. There is wide-ranging evidence, for example, that frequency-derived lexical preferences influence the processing of ambiguity. Reconsider the *Betty knew Monica's date* example, and

replace the verb with *thought*. *Think* is a verb that cannot be followed by a direct object, but only by a clausal complement. Thus *Betty thought Monica's date* would be given a clausal complement analysis, because only a clause such as *had bought flowers* can follow; no ambiguity would be encountered. When verbs can appear in more than one structure (e.g., *admit* can either appear with a direct object or with a clausal complement), empirical research has shown that the structural processor can be biased in its initial analysis towards the more frequent sentences type for this verb (e.g., [14]; but for contradictory results see [30]). More generally, Crocker argues that the pervasive role of experience, as supported by a range of findings revealing the influence of frequency, offers a fundamental explanation for how people adapt to language over time, enabling them to deal so effectively with ambiguity [8]. Both probabilistic [9, 6, 8] and connectionist models of sentence processing [11] (see also Mayberry and Crocker, *The Evolution of a Connectionist Model of Situated Human Language Understanding* of this volume) naturally manifest the central role of experience, favoring interpretations which are supported by evidence they were exposed to during training. As a consequence, experience-based modes fit with a rational view of linguistic performance in which processing mechanisms seek to optimize understanding [5], by recovery of the interpretation most likely to correct, rather than minimize representational or processing complexity [13, 15].

In addition, conceptual knowledge has been shown to influence the processing of syntactic ambiguity. In particular the assignment of thematic roles to noun phrases has served as a test case (e.g., [31, 27]). The thematic roles of a verb describe the mode of participation entities play in the event denoted by the verb: For example, cops usually arrest criminals (and are therefore suitable agents for the event of arresting) whereas criminals usually are being arrested (and are therefore suitable patients for the event of arresting). Reading times in [27], for example, suggest that readers compute and use such event-specific world knowledge immediately for the interpretation of the ambiguous region of reduced relative clauses (*The criminal/cop arrested by* ...) as evidenced by modulation of reading times in the subsequent disambiguation region.

A fourth important factor for resolving structural ambiguity is discourse context. For example, the sentence *Monica told her friend that she had been trying to avoid* ... could be completed with *her date* or with *to call back tomorrow*. In the first case *that she had been trying to avoid* would be an assertion told to Monica's friend; in the later case the same phrase would be a specification for which friend Monica meant. Altmann et al. [1] found that discourse context plays an important part in determining how these sentences are read. For example, if Monica had previously mentioned two friends, then *that she had been trying to avoid* is analyzed by listeners as a distinguishing modification. These findings have also been replicated in situated spoken language comprehension, where the relevant referential context is provided by a visual scene, rather than a prior discourse, crucially highlighting comprehenders' ability to exploit both linguistic and non-linguistic context [35].

This partial survey of psycholinguistic findings, clearly support the notion of a human sentence processing mechanism that is not only incremental but is also highly adaptive to different information sources. Both constraints resulting from

long-term exposure to language, like biases for lexical verb frames or preferences for certain syntactic structures, as well as constraints resulting from short-term exposure, like discourse context, are rapidly exploited as they become available during on-line sentence processing.

1.3 Anticipation in Situated Comprehension

More recently, evidence is mounting that sentence processing is not only incremental, but even anticipatory (e.g., [2]): Listeners are able to rapidly combine information from different constituents to predict subsequent arguments. While the more traditional experimental method of tracking eye movements during reading provided detailed information about the time course of various interpretation processes, anticipatory behavior was not easily detectable with this method. With the advent of eye-tracking in visual scenes [35], however, it became possible to gain clear insight into both the current interpretation listeners are adopting, as well as continuations of a sentence that they expect would plausibly follow from that interpretation. Whereas in reading studies, text is displayed on a computer screen and reading times at different positions in the text (usually the point of disambiguation) allow conclusions about cognitive processing load, in *visual-world* studies participants view scenes depicting objects and events while simultaneously listening to a related utterance. Eye movements are measured in relation to interesting regions of the acoustically presented sentence, such as a noun referring to the objects on the screen. Such utterane-mediated gaze is closely time-locked with the unfolding sentence, with shifts in visual attention occurring about 200 ms after the relevant spoken material is heard. Empirical evidence has further shown that listeners make eye movements in anticipation that a picture in a display will become relevant. For example, upon hearing *the boy will eat*, listeners start looking at edible objects even before they are mentioned [2]. Anticipatory eye movements can thus inform us about higher-level processes, such as the role of verb information in restricting the domain of subsequent reference.

2 Varying Constraints

Outside the laboratory, in the real world, language users have to deal with multiple information sources and modalities simultaneously. Everyday sentences include structural, lexical, discourse, as well as prosodic information in varying degrees; the listeners' task is then to successfully use the relevant information to guide sentence interpretation. It is likely that the impact of different information types changes with varying circumstances; also one information type might be more important than another type, and their impact might happen at different times in the sentence.

Ultimately, any theory of human sentence processing must be able to account for sentence processing in the light of multiple, varying information sources. In responding to this, psycholinguistic research therefore needs to shift away from

simply establishing which information sources influence on-line sentence processing, and place increased emphasis on determining the circumstances under which each type of information source is more or less likely to have an impact. One approach that will bring us closer to achieving this goal is to study comprehension in the face of varying and even contradictory information sources. In this way we explore the extent to which specific information types are favored, dismissed, or weighted with respect to each other. A secondary issue concerns the notion of *task*, as evidence mounts for the view that people process language in importantly different ways depending on whether they are simply reading [32], required to make judgements or answer questions [34], or even to carry out spoken instructions [35]. The exploration and development of such an account of adaptive mechanisms in sentence processing will thus better account for variations in behavior in diverse contexts and tasks.

We present five representative experimental investigations conducted in the context of the ALPHA project that address the issue of sentence processing in light of varying information sources. The first study was concerned with the role of discourse information in word ordering preferences. In this project, we tested whether difficulties with processing non-canonical word orders in German can be weakened with discourse context which provides information about grammatical functions. The second study investigated the interaction of syntactic ordering preferences with prosodic information: In spoken language, intonation contours can convey a range of communicative functions. We tested whether listeners rely on a specific prosodic pattern for the interpretation of German scrambled sentences. The third study looked at the influence of lexical preferences on semantically constrained verb arguments. Semantic verb information is known to restrict listeners' expectations about upcoming verb arguments, and we examined in this study the role of experience with lexical items in forming argument expectations. In the fourth study, the influence of scene objects on linguistic expectations was examined. Whereas in the third study we assessed the long-term constraint of lexical frequency in semantically constraining contexts, in the fourth study we tested the short-term constraint of visual context in semantically constraining utterances. Finally, in the fifth set of experiments, we more deeply investigate the on-line interplay of scene and language processing, and examine the priority of scene information relative to expectations arising from our longer-term world knowledge.

2.1 Discourse Information and Structural Preferences

German is a language with relatively free constituent order. For instance, the initial position in matrix declaratives observes very few restrictions regarding the kind of constituent it can host, which includes subjects, objects, as well as modifiers. Thus both SVO orders like *der Verein St. Johann gewann den Pokal*, "the club$_{NOM}$ St. Johann won the prize$_{ACC}$" and OVS orders like *den Pokal gewann der Verein St. Johann*, "the prize$_{ACC}$ won the club St. Johann$_{NOM}$" are possible in German, though there is clear preference for the canonical subject-first order (see, e.g., [17]). In the

previous example, the nominative case of the subject as well as the accusative case of the object are unambiguously assigned with case marking. However, although German does use morphological case to mark grammatical functions, the system often features syncretism: In many noun phrases (NPs), nominative and accusative cases share surface form. As a result, the constituent ordering of a sentence can be ambiguous: *die Mutter ruft die Tochter*, "the mother$_{NOM,ACC}$ calls the daughter$_{NOM,ACC}$" could either mean that the mother is calling the daughter (SVO) or that the daughter is calling the mother (OVS). In order to correctly interpret an utterance in which the structure cannot be determined on the basis of linguistic information alone, human language users may rely on other information sources to resolve the ambiguity. One such short-term information source might be discourse context. Weber and Neu [40] tested this assumption in a German reading study in which information about grammatical functions of referents could only be inferred from information in the preceding discourse.

In their study, a target sentence with a temporal word order ambiguity was preceded by a question that assigned the grammatical function to one of the referents in the target sentence. In the target sentences, case marking of initial NPs was ambiguous with respect to grammatical function, while case marking of the second NP disambiguated sentences towards SO or OS order (e.g., *Die Katze jagt gleich den Vogel/der Hund mit grossem Eifer*, "the cat$_{NOM,ACC}$ chases in-a-moment the bird$_{ACC}$/the dog$_{NOM}$ with great eagerness"). Without further context, the default interpretation of *die Katze* is subject, since subject-first sentences are the canonical order in German; no processing difficulties should arise upon reading the second object argument *den Vogel*, since it agrees with the subject interpretation of *die Katze*. However, upon encountering a subject as second argument (*der Hund*), readers will have to revise their initial interpretation of *die Katze* as subject; this will be reflected in longer reading times of the second argument *der Hund*. Preceding context consisted of two sentences: a declarative sentence introducing three possible referents (e.g., *Auf der Wiese sind eine Katze, ein Hund und ein Vogel*, "on the field are a cat, a dog and a bird"), followed by a focussing wh-question. Crucially, the focussing question provided information about the grammatical function of a subsequent referent. For instance, the question *Wen jagt gleich die Katze mit grossem Eifer?*, "whom$_{ACC}$ chases in-a-moment the cat with great eagerness?" introduces the cat as subject, the grammatical role the cat will most likely also take in a subsequent answer. The question particles of the focussing questions were either *who* (NOM) or *whom* (ACC). In a baseline condition, a question that did not assign grammatical functions to subsequent NPs was used. For an example of a complete stimulus set see Fig. 1.

There were two reasons for using different question types: For one, both the *who* and the *whom* questions were providing information about the grammatical function of the first NP in the target sentences, whereas the baseline questions did not provide such information. A comparison of focussing questions with the baseline question would therefore inform us whether the processing of sentences with canonical and non-canonical words orders profits from contextual focus. The comparison between *who* and *whom* questions, on the other hand, would inform us about the additional

wh_NOM	Wer jagt gleich den Vogel mit großem Eifer? *Who (NOM) chases in-a-moment the bird with great eagerness?*
wh_ACC	Wen jagt gleich die Katze mit großem Eifer? *Whom (ACC) chases in-a-moment the cat with great eagerness?*
wh_neutr	Was passiert gleich? *What will in-a-moment happen?*
target SO	**Die Katze jagt gleich den Vogel mit großem Eifer.** *The cat (NOM, ambiguous) chases in-a-moment the bird (ACC) with great eagerness.*
wh_NOM	Wer jagt gleich die Katze mit großem Eifer? *Who (NOM) chases in-a-moment the cat with great eagerness?*
wh_ACC	Wen jagt gleich der Hund mit großem Eifer? *Whom (ACC) chases in-a-moment the dog with great eagerness?*
wh_neutr	Was passiert gleich? *What will in-a-moment happen?*
target OS	**Die Katze jagt gleich der Hund mit großem Eifer.** *The cat (ACC, ambiguous) chases in-a-moment the dog (NOM) with great eagernes.*

Fig. 1 Example of stimulus set with three different questions preceding both the SO target sentence and the OS target sentence

influence of structural expectancies; whereas after *who* questions answers are more likely to begin with the subject (SO), after *whom* questions, the object is more likely to be in sentence-initial positions (OS).

Weber and Neu [40] found faster total reading times for the second NP in target sentences (the point of disambiguation) when sentences were preceded by focussing questions (*who* or *whom*) than by the baseline question (see Fig. 2). This supports the assumption that both locally ambiguous canonical and non-canonical word orders

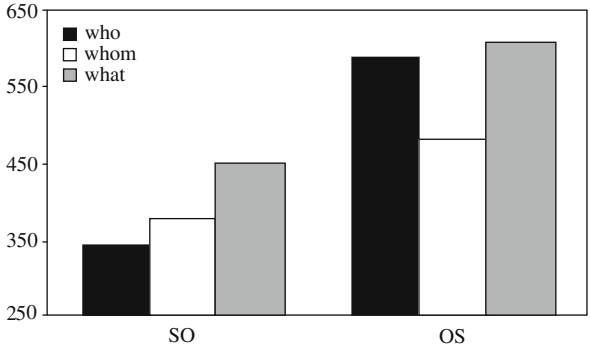


Fig. 2 Total reading times in ms for the second disambiguating NP in SO and OS sentences after a focusing who- or whom-question, and a neutral what-question

profit from focus in a preceding discourse context. Note, however, that OS sentences preceded by focussing questions were still harder to process than comparable SO sentences. So the difficulties of processing non-canonical word orders were not fully overcome by focussing context.

Second, both SO and OS sentences were easier to process when the syntactic structure of the focussing question was matching with the structure of the target sentence. For SO sentences, reading times were faster when the sentence was preceded by a *who* question than by a *whom* question; for OS sentences, reading times were faster when the sentence was preceded by a *whom* question than by a *who* question. This result is in line with findings of syntactic priming in comprehension in which sentences were found to be processed more easily when they were preceded by sentences with a matching syntactic structure than by a mismatching syntactic structure (e.g., [4]). However, when syntactic structure was mismatching, SO sentences in [40] were still easier to process than the baseline condition. Thus, processing can still gain from information about grammatical functions even when there is a structural mismatch. And finally a ray of hope for the processing of non-canonical OS sentences: even though OS sentences were overall more difficult to comprehend than SO sentences, the presence of a focusing question that also matched in syntactic structure resulted at least in reading times that were comparable with the baseline condition of the canonical SO sentences. Thus, short-term information from discourse context can significantly help to overcome processing difficulties with non-canonical word orders in German.

2.2 Prosodic Information and Structural Preferences

A further short-term information source in spoken sentence processing, besides discourse information, is prosody. Prosody is the description of phrasing, stress, loudness, and the placement and nature of pitch accents in spoken language. It can express or aid a range of functions in communication: mark the difference between immediately relevant vs. background information, express contrast, contradiction, correction, or even indicate the intended syntax of ambiguous utterances. Prosody is different from the other information sources in that it is highly variable in its realization. There is, for instance, no simple and direct correspondence between syntactic and prosodic structures. Quite often, a speaker can choose between a number of different intonation contours to express a particular communicative function. Nevertheless, it has been shown that listeners rely on prosodic information in sentence processing. On a structural level, for example, evidence has been presented that prosody can guide listeners' interpretation of attachment ambiguities (e.g., [19]). Sentences with early closure (*When Roger leaves the house is dark*) were compared with late closure sentences (*When Roger leaves the house it's dark*), and using a variety of experimental tasks it was shown that sentences with cooperating prosody (i.e., with a prosodic boundary after *leaves* in the early closure sentence) were processed more quickly than those with baseline prosody. Sentences with conflicting prosody were processed more slowly than those with baseline prosody.

Weber et al. [38] examined the role of prosody in a different ambiguity type, namely word order ambiguity in German. Incorrect initial interpretation of word order typically results in a much stronger garden-path effect than the previously tested modifier attachment ambiguities. One possible reason for this is that reanalysis from an SVO to an OVS structure entails a complete reassignment of the verbs' roles to both arguments. Given the stronger-garden path effect, it is particularly interesting to attest the role of prosody in this ambiguity type.

As described in the previous section, German nominative and accusative case often share surface forms. In combination with free constituent order in German, a functional gap arises: for example, *die Katze*, "the cat", in utterance initial position can be both subject (nominative case) and object (accusative case). In an eye-tracking study with visual scenes, Weber et al. [38] examined whether prosody can fill the functional gap arising from a combination of syncretism and free constituent order in German. Can prosody, in the absence of unambiguous morphological and configurational information, influence the assignment of grammatical function?

To investigate this question, they observed anticipatory eye movements of German listeners in a scene during comprehension of a related utterance. Not only has it repeatedly been shown that referents in a scene are identified as soon as they are referred to in an utterance, there are several studies revealing that they can be identified prior to their mention. With respect to constituent order ambiguity in German, two eye-tracking studies priorly attested such anticipatory behavior. For one, Kamide et al. [18] have shown that unambiguous case marking, combined with verb selectional information, leads to post-verbal anticipatory eye movements in German SVO and OVS sentences. That is, upon hearing *der Hase frisst...*, "the hare$_{NOM}$ eats...", German participants start to look at an appropriate object argument in the scene (e.g., a cabbage) even before hearing the second argument; upon hearing *den Hasen frisst...*, "the hare$_{ACC}$ eats...", they anticipate an appropriate subject argument (e.g., a fox). Thus, listeners are able to use case marking to assign the appropriate grammatical function to the first acrgument and combine this with the semantics of the verb, resulting in increased anticipatory fixations to the appropriate second argument.

Weber et al. [38] similarly employed German SVO and OVS structures, but with sentence-initial NPs that were ambiguously marked for nominative or accusative case. Morphosyntactic disambiguation of grammatical functions took place at the second NP that was clearly case marked as either accusative or nominative (e.g., *Die Katze jagt womöglich den Vogel/der Hund*, "the cat$_{NOM,ACC}$ chases possibly the bird$_{ACC}$/the dog$_{NOM}$"). Scenes accompanying the sentences showed the referent of the first NP (e.g., a cat) and plausible objects and subjects for the referent of the first NP in relation to a given action (e.g., a bird as plausible object for being chased by a cat and a dog as plausible subject for chasing a cat, see Fig. 3). No actions were depicted. Thus, even though the scenes presented potential referents they could not help with disambiguating grammatical roles in any way (see Sect. 2.5). In contrast with previous studies, however, prosodic cues could potentially help listeners resolve the temporary SVO/OVS ambiguity.

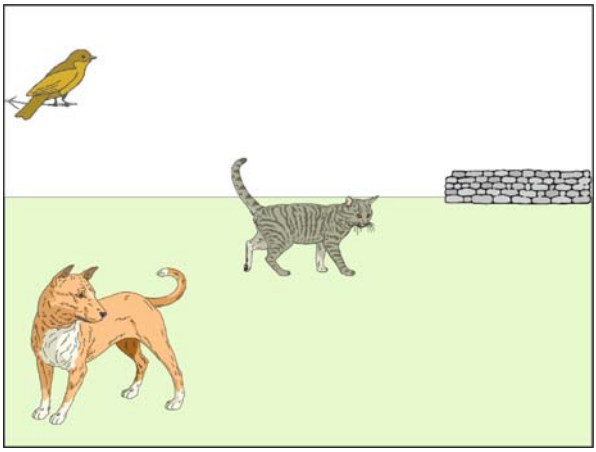


Fig. 3 Visual context for spoken sentences: *Die Katze jagt womöglich den Vogel/derHund*

The SVO sentences had a low pitch accent on the first NP (L* + H according to GToBI transcription [16]), followed by a focal high-pitch accent (H*) on the verb. This prosodic pattern was considered unmarked and was expected to indicate canonical subject-first sentences. The OVS sentences had a focal high-pitch accent (L + H*) on the first NP. This prosodic pattern was considered marked and was expected to indicate non-canonical object-first sentences. During the verb (e.g., chases), no effect of prosody was found. That is, potential objects (e.g., a bird) were fixated more often than potential agents (e.g., a dog) in both SVO and OVS sentences. Looks to the potential object imply that the initial NP was interpreted as subject (and therefore agent). The preference for anticipatory looks to potential objects at this point is a mere reflection for the well-attested preference for the canonical SVO order in German. During the following adverb (e.g., possibly), however, only for SVO structure more looks to potential objects were found. For OVS structures, potential subjects drew slightly more looks than potential objects. Thus, the strong preference for an SVO interpretation disappeared in sentences with OVS-type intonation. Prosodic cues were interpreted rapidly enough to affect listeners' interpretation of grammatical function before disambiguating case information was available.

The influence of prosodic information on the resolution of word order ambiguity is particularly striking for two reasons. First, the preference for the canonical SVO structure is very strong for German listeners. This is not surprising given that only about 18% of German sentences are OVS (in the Negra corpus; [39]). Most likely, this preference is stronger than that of previously tested attachment ambiguities. Prosodic information is therefore competing against a structural preference that has found plenty of support from long-term exposure to language and is highly ingrained. Second, as mentioned before, prosodic realizations are variable. A nuclear pitch accent on the first NP is definitely not the only way to intone an OVS structure. Intonation contours with phrase breaks or silent intervals after the first NP are also easily imaginable, for example. In addition, a nuclear pitch accent

on the first NP can have in a different context a different meaning; for instance, the same pitch accent is known to convey contrasts. Further research is necessary to test whether other prosodic patterns can similarly influence the interpretation of grammatical functions. For the specific situation of the described study, however, we could show that prosodic focus on the first NP in OVS sentences placed a high prominence on the noun phrase which in turn facilitated the interpretation of the marked syntactic structure OVS.

2.3 Semantic Information and Lexical Preferences

Similar to the anticipation of arguments based on grammatical information we described above, anticipatory behavior in eye-tracking studies has been found for semantically constraining verb information. That is, listeners start looking at pictures of suitable object NPs right after semantically constraining verbs [2]: following *the boy will eat*, listeners fixate edible objects in a scene even before they are mentioned in the utterance. The semantic information extracted at the verb is sufficient to exclude other visually presented objects as potential referents. This entails that the human processor can immediately establish anaphoric dependencies on the basis of thematic fit between the referents in the visual context and the verb.

At the same time, there is ample evidence that the human processor has lexical biases which are built on long-term experience; words that occur more often in a language are favored and recognized more easily than less frequent words (e.g., [25]). In particular, the simultaneous activation of word candidates with overlapping onset has been shown to be modulated by lexical frequency [10]: While hearing the word *bench*, English listeners look more at the distractor picture of the high-frequency *bed* than at the distractor picture of the low-frequency *bell* in an eye-tracking study. The combination of semantic information and lexical preferences can lead to a situation in which verb information constrains potential referents in the presence of semantically inapt high-frequency distractors. Weber and Crocker [36] investigated the interaction of lexical frequency effects with effects from verb constraints in a German eye-tracking study with visual scenes. In particular, they tested whether high-frequency distractors are activated even though semantic information from preceding verbs renders them unlikely word candidates.

In their study, German participants listened to sentences with restrictive and unrestrictive verbs (e.g., *Die Frau bügelt/sieht die Bluse*, "the woman is ironing/seeing the blouse") while they were looking at a display with four objects. The display showed the agent of the sentence (e.g., *Frau*, "woman"), a low frequency target (e.g., *Bluse*, "blouse"), a high-frequency phonological distractor (e.g., *Blume*, "flower"), and an unrelated distractor (e.g., *Wolke*, "cloud"; high in lexical frequency but phonologically unrelated to the target) (see Fig. 4). From the view of semantic information, the target and the distractors are possible object arguments following the unrestrictive verb (e.g., is seeing), but only the target is a likely candidate following the restrictive verb (e.g., is ironing). From the view of lexical frequency, however,

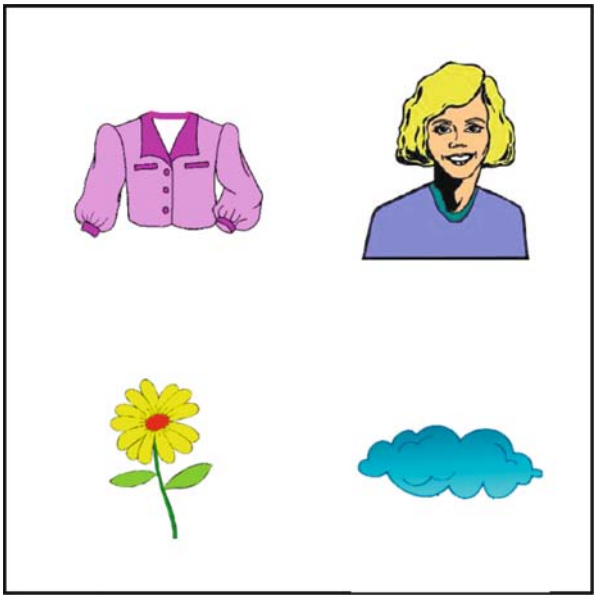


Fig. 4 Visual context for spoken sentences: *Die Frau bügelt/sieht die Bluse*

the high-frequency phonological distractor *Blume* should draw more looks than the low frequency target while hearing the ambiguous part of the target (e.g., /blu/ in "Bluse").

Not surprisingly, Weber and Crocker [36] replicated the finding that when the verb was not semantically constraining the set of potential object arguments, German listeners fixated both the picture of the target and the picture of the phonological distractor during the ambiguous part of the target. Thus, both *Bluse* and *Blume* were considered as potential object arguments following the unrestrictive verb *siehst*. No activation of the phonological distractor was, however, observed when the preceding verb was excluding the distractor as a likely object referent (see Fig. 5); looks went almost exclusively to the target *Bluse* following the restrictive verb *bügelt*. At first glance, it clearly seems that semantic information provided by the verb is sufficient to exclude semantically inappropriate distractors even when they are high in lexical frequency. However, this complete lack of distractor activation in semantically constraining context should be taken with some caution; lexical frequency was predicted after all to make the phonological distractor more attractive than the target, albeit only when there is no semantic restriction on the target. This was, however, not what Weber and Crocker [36] found; rather the picture of the target and the phonological distractor were equally attractive in unrestrictive sentences. This seems surprising given the earlier findings of lexical frequency effects in eye tracking (see [10]).

In contrast to these earlier studies, the German participants in [36] had no specific task during the experiment, other than to listen to the speech and to look at

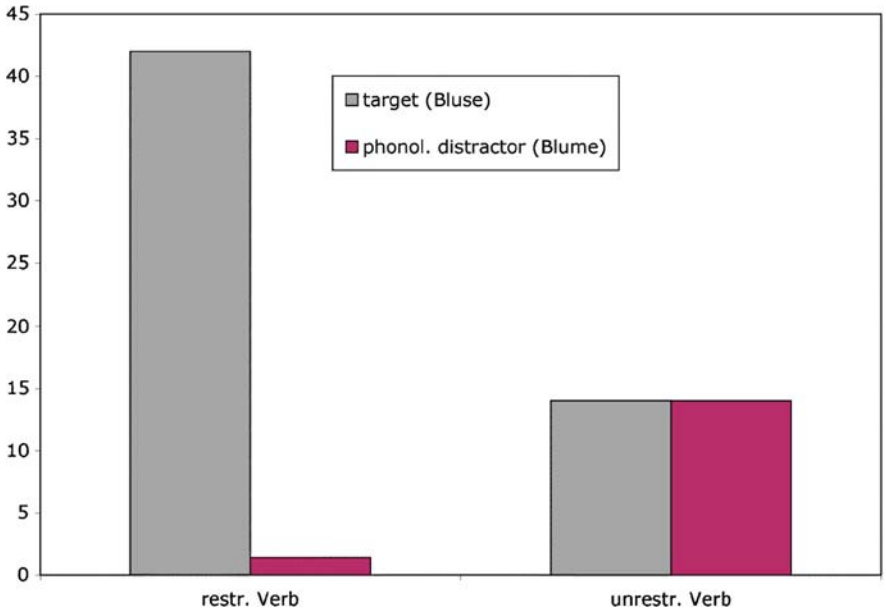


Fig. 5 Attractiveness of target and phonological distractor between 300 and 600 ms after target onset, measured as added percentages of looks over unrelated distractor. No specific task

the screen. Previously, targets in lexical frequency studies had been presented in semantically empty carrier phrases which simply instructed participants to click on a displayed object (e.g., click on the blouse). In a second experiment, Weber and Crocker [36] therefore tested whether, in combination with a task, lexical frequency effects could be observed with their materials. They presented the same materials to a new set of German listeners, the only difference being that participants were told to click on the picture of the second argument in the sentence. This time, the phonological distractor *Blume* was indeed more attractive than the target *Bluse* in unrestrictive sentences (see Fig. 6). Just by having an explicit task, lexical frequency effects emerged. This dominance of high-frequency phonological distractors is therefore consistent with previous studies on lexical frequency effects that employ a click task.

But even more interesting for the question of semantic information, Weber and Crocker [36] found activation of the phonological distractor in semantically constraining contexts; even though the verb information in *bügelt* should have rendered the phonological distractor *Blume* an unlikely candidate for the object argument, German listeners still look at it more than would be expected. In the constraining sentences, the target was overall more attractive than the phonological distractor, but the phonological distractors drew also a considerable proportion of looks. This suggests that effects of preceding verb information can indeed be modulated by lexical frequency.

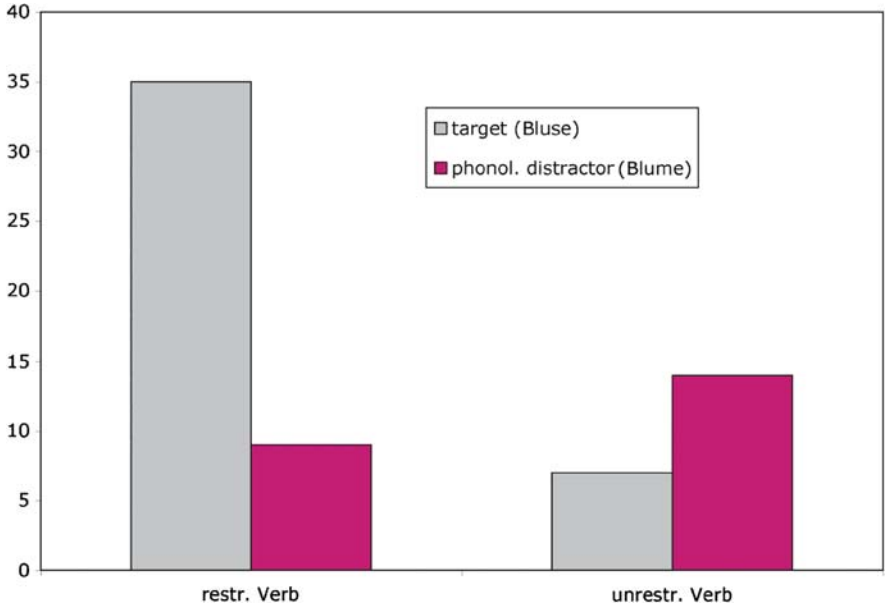


Fig. 6 Attractiveness of target and phonological distractor between 300 and 600 ms after target onset, measured as added percentages of looks over unrelated distractor. Clicking task

The fact that activation of semantically inappropriate, high-frequency distractors was only found when the participants' task was to click on the last argument in the sentence, suggests that frequency effects in eye tracking are sensitive to task specific demands. Apparently, only when listeners' attention is purposefully directed to the verb arguments, are frequency effects observable. This finding speaks for a human parser that is not only applying different information sources incrementally, but that is also sensitive to cognitive task demands.

2.4 Semantic Information and Visual Context

We have observed above that situated language comprehension rapidly directs visual attention in a relevant scene, both to mentioned and anticipated referents. An important question about these findings is the extent to which they are indicative of general comprehension mechanisms, or whether the scene objects themselves contribute to the forming of specific expectations for verb arguments. Weber and Crocker [37] therefore investigated further the influence of visual context on constraining verb information in a cross-modal priming experiment.

Lexical decision times are known to be faster following semantically related objects than semantically unrelated objects; that is, listeners respond faster to *nurse* after *doctor* than after *grass* (e.g., [28]). Also verbs have been shown to prime typical agents, patients, and instruments (e.g., [12]). In a first step, Weber and Crocker

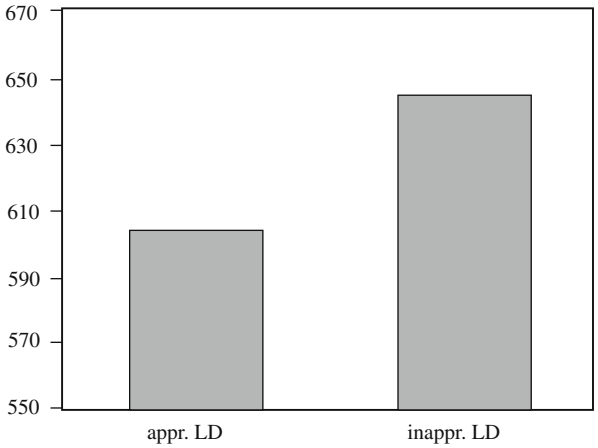


Fig. 7 Average lexical decision times for semantically appropriate and inappropriate object arguments. Only auditory prime

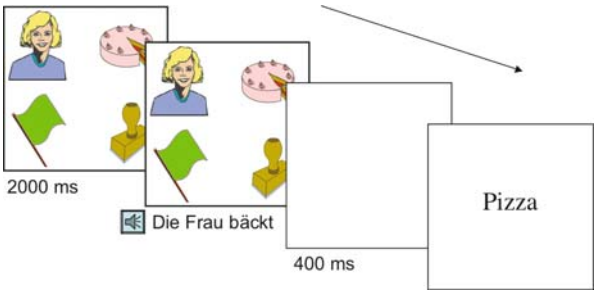


Fig. 8 Example for a trial with a combination of auditory and visual primes

[37] replicated this finding for German with a simple cross-modal priming study in which selectional verb information could prime object arguments. In their study, German listeners were presented with sentence onsets which had a semantically restrictive verb (e.g., *Die Frau bäckt*, "the woman bakes"); a lexical decision task for visually presented nouns followed the auditory prime sentence fragment. The visual lexical decision items were either semantically appropriate as arguments for the verb (e.g., *Pizza*, "pizza") or inappropriate (e.g., *Palme*, "palm tree") as arguments for the verb. As expected, reaction times were faster for semantically appropriate items than for inappropriate ones (see Fig. 7), replicating the well-known semantic priming effect for German.

In order to further investigate the influence of the visual context on forming expectations about upcoming verb arguments, Weber and Crocker displayed in a second study objects on a screen, simultaneously with the auditory prime (see Fig. 8). The displays were typical for eye-tracking studies and showed four objects: the agent of the sentence onset (e.g., *die Frau*, "the woman"), an object either

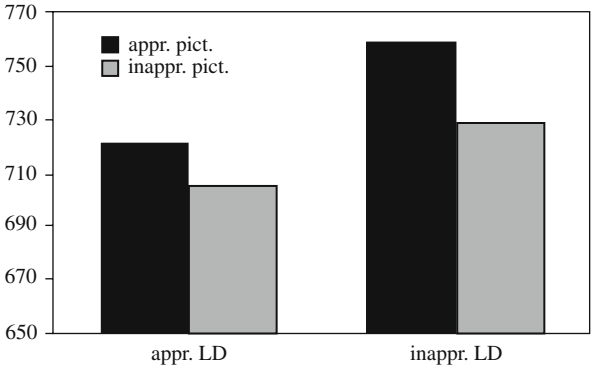


Fig. 9 Average lexical decision times for semantically appropriate and inappropriate object arguments. Auditory and visual prime

semantically appropriate as argument for the verb (e.g., *Torte*, "pie") or inappropriate (e.g., *Tanne*, "pine") and two distractor objects.

As before, a lexical decision task to visually presented nouns followed the primes. Both the appropriate visual argument (e.g., *Torte*, "pie") and the appropriate lexical decision item (e.g., *Pizza*, "pizza") were highly plausible arguments for the sentence onsets (as defined by a rating study). As in the first study, reaction times were faster for lexical decision items which were semantically appropriate than for items which were inappropriate (see Fig. 9). Surprisingly, however, reaction times were slowed when the display included a picture of an appropriate argument prior to lexical decision. This semantic interference (rather than facilitation) from appropriate pictures occurred both when lexical decision items were appropriate and when they were inappropriate. Thus, visual context did influence reaction times in the sense that it gave competition to the lexical decision items. On the other hand, facilitated lexical decision times for appropriate items, regardless of the scene, provide evidence for a purely linguistic anticipation of upcoming verb arguments (confirming the gated completion findings of [2]). We suggest that visually attending the picture of an appropriate object based on supporting auditory input leads to contextually grounded expectations concerning which object would follow as the verb argument; when the visually supported expectations were not met by the target word, lexical decision times were slowed across the board.

2.5 The Influence of the Scene: Depicted Events and Their Priority

The studies described above provide diverse evidence supporting the notion that the human language comprehension system is able to rapidly adapt to, and exploit, a range of linguistic information sources: discourse context, prosody, lexical frequency, and verb semantics. We further noted that anticipatory inspection of relevant depicted referents not only reflects incremental interpretation and expectations, but

further instantiates those expectations with the depicted object. A natural question in the case of situated language processing, therefore, is whether more complex scene information can influence spoken language understanding. Previous work by Tanenhaus and colleagues [35] has shown, for example, the rapid influence of visual referential context on ambiguity resolution in on-line situated utterance processing. Listeners were presented with a scene showing either a single apple or two apples, and the utterance *Put the apple on the towel in the box*. Eye-movements revealed that the interpretation of the phrase *on the towel*, as either the location of the apple versus its desired destination was influenced by the visual context manipulation. Sedivy et al. [33] further demonstrated the influence of a visual referential contrast: listeners looked at a target referent (e.g. the tall glass) more quickly when the visual context displayed a contrasting object of the same category (a small glass) than when it did not.

An eye-tracking study by Knoeferle et al. [21] investigated the interpretation of German SVO and OVS sentences with case-ambiguous initial NPs. Structural disambiguation took place only at a second NP that was clearly case marked as either nominative or accusative (e.g., *die Prinzessin malt offensichtlich den Fechter/der Pirat*, "the princess$_{NOM,ACC}$ paints apparently the fencer$_{ACC}$/the pirat$_{NOM}$"). In the accompanying scenes, however, depicted actions were potentially able to resolve the ambiguity as soon as the verb was encountered (see Fig. 10). Their findings revealed anticipatory post-verbal eye movements to the appropriate second argument based on verb-mediated identification of the relevant scene event, and crucially before the disambiguating second NP was heard. The time-course and pattern of gaze clearly suggest that listeners were able to use depicted events to resolve the ambiguity and assign grammatical functions appropriately, just as they have been shown to use linguistic [18] and prosodic [38] constraints.

Given that information sources as diverse as syntax, semantic, intonation, and depicted events can so rapidly and effectively be used during situated spoken

Fig. 10 Visual context for spoken sentences: *Die Princessin wäscht/malt gerade* ...

language comprehension, Knoeferle and Crocker [20] investigated the time course with which world knowledge about typical events [2] and information from the atypical scene events interacted [21], and, crucially, the *relative importance* of these information sources. In a German eye-tracking study, they investigated the anticipation of both stereotypical role-fillers, based on verb expectations, and depicted role-fillers, based on depicted events in syntactically unambiguous sentences.

Their findings confirmed, in a single study, that people are able to rapidly and equally exploit both information sources, linguistic or visual, when either kind of constraining information is available to anticipate thematic role fillers. Crucially, however, when they pitted the two information sources against each other, they observed a greater relative importance of verb-mediated depicted events information over stereotypical thematic role knowledge associate with the verb. When listeners heard a sentence beginning *Den Pilot bespitzelt gleich* . . . , the verb (e.g., *bespitzelt*) (spies-on) identifies two different agents on a scene as relevant, participants prefer upcoming agents that match with a displayed action (e.g., *a wizard*) over agents that match with their world knowledge (e.g., *a spy*) (see Fig. 11). Eye movements to the agent depicting the action of the verb occur shortly after the verb and crucially before the agent was mentioned in the utterance.

To further investigate the priority and use of scene events, Knoeferle and Crocker [22] conducted a series of experiments investigating the temporal interdependency between *dynamic* visual context and utterance comprehension. Exploiting the "blank screen paradigm", event scenes were presented prior to the onset of an utterance and then replaced by a blank screen either before or during the utterance. Additionally, two of the experiments featured scenes involving dynamic events, i.e., actions were depicted as occurring over time, introducing an aspectual dimension to the depicted events, which were furthermore coupled with verb and adverb tense manipulations in the utterances used in the third experiment. The findings suggested that people do use scene event information even when it is no longer

Fig. 11 Visual context for spoken sentences: *Den Pilot bespitzelt* . . .

present, but that the relative priority with respect to other information sources is strongest when events are co-present, and may decay over time.

To account for both the rapid interaction of linguistic and visual information and the observed preference for the information from depicted events, Knoeferle and Crocker [20] posit the Coordinated Interplay Account (CIA), which outlines how the unfolding utterance guides attention in the visual scene to establish reference to objects and events. Once these are identified, the attended information rapidly influences comprehension of the utterance, allowing the anticipation of upcoming arguments not yet mentioned by virtue of their relationship to the objects and events thus established. The close temporal interaction between comprehension and attention in the scene is suggested as the principal reason for the relative priority of the immediately depicted information over stereotypical knowledge in situated comprehension. Knoeferle and Crocker [20] conjecture that there may be a development basis for this preference, arising from the important role that the immediate environment plays as a child learns to ground concepts with visual referents during language acquisition. Mayberry et al. [26] have furthermore developed a connectionist model which instantiates the CIA. The architecture, described in detail in the Chapter by Mayberry and Crocker (this volume), models many of the findings described above, including the priority of scene event information.

In more recent work, Knoeferle and Crocker [22] further refine the CIA to incorporate a working memory (WM) component that contains the current interpretation of an utterance, expectations based on linguistic and world knowledge, and information from objects and events in a dynamic scene (Fig. 12). In order to explain the reduced priority of events that are no longer co-present, the account postulates that

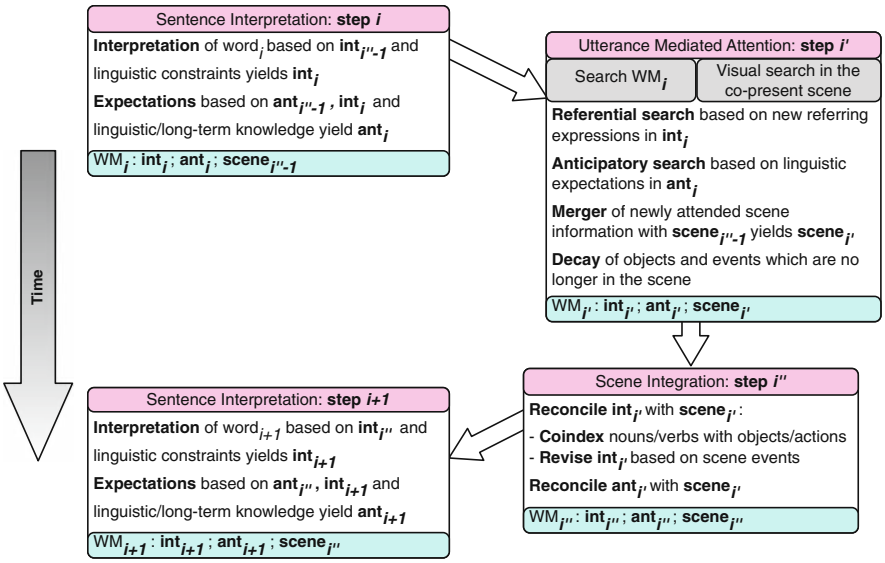


Fig. 12 The Coordinated Interplay Account (CIA): The time course of processes, informational dependencies, and working memory in situated spoken language comprehension [22]

items in working memory decay with time, affecting their influence on the developing interpretation of the unfolding utterance. The introduction of working memory into the CIA in which the accessibility of representations of scene objects and events are dependent on their decaying activation provides a reasonable explanation for the observed effects that is also in accord with current theories of sentence comprehension (see Lewis and Vasishth [23] for discussion as well as [3] for a broader view of the role of working memory in cognition).

3 Conclusions

Human sentence processing is not only incremental but also anticipatory: upon encountering the initial words of a sentence, not only do people immediately begin constructing detailed interpretations, they also initiate hypotheses or expectations about what is likely to follow. The empirical research of the ALPHA project focussed mainly on two aspects of such anticipatory behavior. In the first phase of the project the emphasis of empirical research lay in establishing the information sources which contribute to incremental interpretation and anticipation of forthcoming arguments. It was found that such sources include morphosyntactic and lexical verb information (e.g., [18]), world-knowledge (e.g., [29]), and information from the visual context (e.g., [21]). In the second phase of the project, the focus of empirical research was to determine the extent to which initial interpretation preferences are influenced by different short-term and long-term constraints such as lexical frequency, linguistic context, visual context, and prosodic information. This chapter has highlighted some of the most important empirical findings from the second phase.

While long-term exposure to language can result, for instance, in preferences for certain syntactic structures, biases for lexical verb frames, and frequency effects for lexical choices, recent linguistic and visual context is also exploited on-line to influence understanding. Given the diverse nature of long-term and short-term constraints it seems possible that they affect the comprehension processes differently: for instance, one type of constraint could be dominant. It would, for instance, appear plausible that long-term knowledge derived from experience with language and the world is always accorded greater weight than short-term constraints. Long-term constraints are presumably *routinized* within the processing mechanisms, while short-term constraints may be more variable depending on the specifics of the communicative situation and task. Alternatively, the *here and now* relevance of a communicative situation may foreground short-term constraints in the immediate (linguistic and non-linguistic) context over what we know based on our long-term experience. The first account is appealing since rapid and preferred reliance on long-term experience would enable efficient processing because such long-term knowledge is readily available from memory. On the other hand, an ever-changing dynamic environment and the necessity of adapting to different communicative situations and tasks would appear to favor the second account, placing emphasis on the use of short-term contextual constraints.

Consider our findings in the light of these two accounts. On the one hand, they confirm that long-term biases (e.g., structural bias towards SVO) cannot be fully overridden by short-term contextual constraints that are linguistic in nature: While both information about grammatical function in preceding context (see Sect. 2.1) and prosodic marking of object-first sentence (see Sect. 2.2) could weaken the processing difficulties usually encountered with object-first structure, these short-term constraints were not sufficient to fully generate the interpretation of an object-first sentence. Similarly, lexical frequency could modulate, but definitely not fully change the expectations for an object argument based on restrictive verb information (see Sect. 2.3), and also the results from Sect. 2.4 speak clearly for an interplay of both the visual context and the verb information.

On the other hand, short-term constraints arising from depicted event information appear to dominate (and not just modulate) the stereotypical knowledge of the actions an agent performs (see [20]). This finding together with the strong influence of depicted events on structural disambiguation of locally structurally ambiguous German utterances (see [21]) suggests an account of situated language comprehension in the tradition of the Coordinated Interplay Account posited by Knoeferle and Crocker [20, 22]. In situated comprehension situations, information from the immediate visual context, at least when it depicts role relations between event participants, is accorded great importance for on-line language comprehension.

An added dimension with respect to the use of short- and long-term constraints comes from the presented effects of task: lexical frequency only biased the anticipation of a visually presented object when the task was to click on the target object (Sect. 2.3). The studies in Sect. 2.4 further revealed an interesting combination of long- and short-term constraints: While we observed clear support for the general anticipation of objects based on verb-derived expectations, the scene then instantiated these expectations causing interference with the lexical decision targets that did not match objects in the scene when these general expectations identified a plausible referent in the scene. The pattern of observations is consistent with the Coordinated Interplay Account, which generally argues for the influence of scene information once it has been identified by the utterance as relevant, typically through explicit or anticipated reference to scene objects or events. Taken together, the empirical findings of the ALPHA project speak for a human sentence comprehension system that rapidly integrates diverse informational constraints, derived from both long-term experience and the immediate context, and weighs them depending on the situation and task.

References

1. Altmann, G., Garnham, A., Dennis, Y. Avoiding the garden path: Eye movements in context. Journal of Memory and Language, 31:685–712 (1992).
2. Altmann, G.T.M., Kamide, Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73:247–264 (1999).
3. Baddeley, A.D. Working Memory. Oxford, UK; New York: Oxford University Press (1986).

4. Branigan, H., Pickering, M., Liversedge, S. Syntactic priming: investigating the mental representation of language. Journal of Psycholinguistic Research, 24:489–506 (1995).
5. Chater, N., Crocker, M.W., Pickering, M. The rational analysis of inquiry: The case for parsing. In N. Chater, M. Oaksford (Eds.), Rational Analysis of Cognition (pp. 441–468). Oxford, UK; New York: Oxford University Press (1998).
6. Crocker, M., Keller, F. Probabilistic grammars as models of gradience in language processing. In G. Fanselow et al. (Ed.), Gradience in Grammar: Generative Perspectives (pp. 227–245). Oxford, UK; New York: Oxford University Press (2006).
7. Crocker, M.W. Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language. Dordrecht: Kluwer (1996).
8. Crocker, M.W. Rational models of comprehension: Addressing the performance paradox. In A. Cutler (Ed.), Twenty-First Century Psycholinguistics: Four Cornerstones (pp. 363–380). Hillsdale, NJ: Lawrence Erlbaum Associates (2005).
9. Crocker, M.W., Brants, T. Wide-coverage probabilistic sentence processing. Journal of Psycholinguistic Research, 29(6):647–669 (2000).
10. Dahan, D., Magnuson, J., Tanenhaus, M. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology, 42:361–367 (2001).
11. Elman, J.L. Finding structure in time. Cognition Science, 14(2):179–211 (1990).
12. Ferretti, T., McRae, K., Hatherell, A. Integrating verbs, situation schemas, and thematic role concepts. Journal of Memory and Language, 44:516–547 (2001).
13. Frazier, L., Fodor, J. The sausage machine: A new two-stage parsing model. Cognition, 6:291–325 (1978).
14. Garnsey, S., Pearlmutter, N., Myers, E., Lotocky, M. The contribution of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language, 37:58–93 (1997).
15. Gibson, E. Linguistic complexity: Locality of syntactic dependencies. Cognition, 68:1–76 (1998).
16. Grice, M., Baumann, S. Deutsche intonation und GToBI. Linguistische Berichte, 191: 267–298 (2002).
17. Hemforth, B. Kognitives Parsing: Repräsentation und Verarbeitung Sprachlichen Wissens. Sankt Augustin: Infix-Verlag (1993).
18. Kamide, Y., Scheepers, C., Altmann, G. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. Journal of Psycholinguistic Research, 32:37–55 (2003).
19. Kjeelgaard, M., Speer, S. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. Journal of Memory and Language, 40:153–194 (1999).
20. Knoeferle, P., Crocker, M. The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. Cognitive Science, 30:481–529 (2006).
21. Knoeferle, P., Crocker, M., Scheepers, C., Pickering, M. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye movements in depicted events. Cognition, 95:95–127 (2005).
22. Knoeferle, P., Crocker, M.W. The influence of recent scene events on spoken comprehension: Evidence from eye movements. Journal of Memory and Language (Special Issue on Language-Vision Interaction), 57(2):519–543 (2007).
23. Lewis, R.L., Vasishth, S., Dyke, J.A.V. Computational principles of working memory in sentence comprehension. Trends in Cognitive Science, 10:447–454 (2006).
24. Marslen-Wilson, W. Linguistic structure and speech shadowing at very short latencies. Nature, 244:522–523 (1973).
25. Marslen-Wilson, W. Activation, competition, and frequency in lexical access. In G. Altmann (ed.), Cognitive Models of Speech Processing (pp. 148–172). Cambridge, MA: MIT Press (1990).
26. Mayberry, M., Crocker, M., Knoeferle, P.A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In 28th Annual Conference of the Cognitive Science Society. Vancouver, Canada (2006).

27. McRae, K., Spivey-Knowlton, M., Tanenhaus, M. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language, 38:283–312 (1998).
28. Meyer, D., Schvaneveldt, R. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90:227–234 (1971).
29. Muckel, S., Scheepers, C., Crocker, M., Muller, K. Anticipating German particle verb meanings: Effects of lexical frequency and plausibility. In 8th Annual Conference on Architectures and Mechanisms for Language Processing. Tenerife, Spain (2002).
30. Pickering, M., Traxler, M., Crocker, M. Ambiguity resolution in sentence processing: Evidence against likelihood. Journal of Memory and Language, 43:447–475 (2000).
31. Rayner, K., Carlson, M., Frazier, L. The interaction of syntax and semantics during sentence processing. Journal of Verbal Learning and Verbal Behavior, 22:358–374 (1983).
32. Rayner, K., Raney, G.E. Eye movement control in reading and visual search effects of word frequency. Psychonomic Bulletin & Review, 3:245–248 (1996).
33. Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., Carlson, G.N. Achieving incremental semantic interpretation through contextual representation. Cognition, 71:109–148 (1999).
34. Swets, B., Desmet, T., Clifton, C., Ferreira, F. Underspecification of syntactic ambiguities: Evidence from self-paced reading. Memory & Cognition, 36:201–216 (2008).
35. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C. Integration of visual and linguistic information in spoken language comprehension. Science, 268:1632–1634 (1995).
36. Weber, A., Crocker, M. Top-down anticipation versus bottom-up lexical access: Which dominates eye movements in visual scenes? In 19th Annual CUNY Conference On Human Sentence Processing. New York City, New York (2006).
37. Weber, A., Crocker, M. The influence of the scene on linguistic expectations: Evidence from cross-modal priming in visual worlds. In 20th Annual CUNY Conference On Human Sentence Processing. New York City, New York (2007).
38. Weber, A., Grice, M., Crocker, M. The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. Cognition, 99:B63–B72 (2006).
39. Weber, A., Müller, K. Word order variation in German main clauses: A corpus analysis. In Proceedings of the 20th International Conference on Computational Linguistics (pp. 71–77). Geneva (2004).
40. Weber, A., Neu, J. Assignment of grammatical functions in discourse context and word-order ambiguity resolution. In 16th Annual CUNY Conference On Human Sentence Processing. Boston, Massachusetts (2003).

The Evolution of a Connectionist Model of Situated Human Language Understanding

Marshall R. Mayberry and Matthew W. Crocker

1 Introduction

The Adaptive Mechanisms in Human Language Processing (ALPHA) project features both experimental and computational tracks designed to complement each other in the investigation of the cognitive mechanisms that underlie situated human utterance processing. The models developed in the computational track replicate results obtained in the experimental track and, in turn, suggest further experiments by virtue of behavior that arises as a by-product of their operation. The experiments conducted in the ALPHA project have built upon over a decade of psycholinguistic research in the *visual worlds paradigm* to investigate the interaction between language and visual context. In earlier visual world studies, participants' gazes in a visual scene are monitored while they listen to an utterance, the analysis of which reveals the integration of what they hear, what they see, and what they know [5, 18]. Analysis of eye movements as an unfolding utterance is processed has been successfully used to investigate language understanding at different levels of processing. Findings from the visual world studies have revealed the rapid and incremental influence of visual referential context [18, 17]. Further research demonstrated that listeners even actively hypothesize and anticipate likely upcoming role fillers in the scene based on their general knowledge [1, 9, 8].

Recent research in our group also has shown that depicted events [11] influence the resolution of structural ambiguity in online situated utterance processing. The investigation of the time course with which both linguistic knowledge and scene information are used, in order to establish the relative priority of these different information sources, suggested that while both linguistic knowledge and scene interpretation are used equally well and quickly, the information from the scene has priority when the two sources conflict [10].

These findings led to the coordinated interplay account (CIA; [10]) that took into consideration the role that attention plays in the interaction between the visual

M.R. Mayberry (✉)
School of Social Sciences, Humanities and Arts, UC Merced, CA, USA
e-mail: marty.mayberry@gmail.com

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_8,

context and language and how different information sources are prioritized. The CIA highlights the following *Cognitive Characteristics* of situated spoken language comprehension:

1. *Incremental*: Utterance interpretation is developed as each word is processed.
2. *Anticipatory*: Likely continuations are inferred from current interpretation.
3. *Integrative*: Multiple information sources bear directly on comprehension.
4. *Adaptive*: Comprehension exploits relevant information when it is available.
5. *Coordinated*: Sources of information may temporally depend on each other.

In this chapter, we describe two connectionist models of situated utterance comprehension that yield insights into these Cognitive Characteristics. Connectionist models are an abstraction of the structure of the human brain: computation in these systems occurs in parallel over massively interconnected simple processing units (neurons). The non-modular nature of connectionist systems makes them well-suited to modeling the cognitive characteristics listed above because information that the system has learned is distributed throughout its weights (i.e., the connections from one unit to another). Thus information distribution leads to a number of useful properties: pattern completion, fault tolerance, generalization, gradience, and prototype formation, just to mention a few. Such systems function on the basis of constraint satisfaction to exhibit behavior sensitive to the dynamic statistics of the five cognitive characteristics in a manner remarkably similar to people. The earlier of the two models we describe, which we will dub EVTNET for this chapter, used compressed *event layers* to represent the scene events distinctly [12]. EVTNET was primarily motivated by these cognitive characteristics, but did not model attention. A refined model, CIANET [13], was developed that incorporated an explicit attentional mechanism to dynamically select and bind constituents of the event from the scene that is most relevant to the unfolding interpretation as it is being processed. This mechanism enables CIANET to go beyond merely fitting data to predicting how people resolve conflicting information when only exposed to non-conflicting training exemplars. CIANET demonstrates four *Modeling Goals* that allow the system to instantiate the CIA: (1) it exhibits the Cognitive Characteristics of human utterance comprehension noted above; (2) it employs an *explicit* attentional mechanism that gives rise to this cognitively plausible behavior; (3) it models the *empirically* observed preference for depicted information over stereotypical knowledge depending on experience; and (4) it provides support for a *developmental* basis of this preference in the capacity of a connectionist system in which learning is central to how the system operates.

2 Experimental Findings

The experiments we model were conducted in German, a language that has a number of features that make it ideal for studying incremental thematic role-assignment. In German, both subject-verb-object (SVO) and object-verb-subject (OVS) ordering

are grammatical, though SVO word order is canonical. The case-marking on the determiner in a noun phrase typically indicates the grammatical function of the noun phrase. However, for feminine and neuter nouns, the nominative and accusative case markings on the determiner is identical. The flexibility in word order together with the case ambiguity results in local structural ambiguity and permits the investigation of how syntactic constraints, world knowledge, and scene information contribute to the resolution of structural and role-assignment ambiguity.

2.1 Anticipation in Unambiguous Utterances

The first two experiments modeled involved unambiguous utterances in which case-marking and verb selectional restrictions in the linguistic input, together with depicted characters in a visual scene, allowed rapid assignment of the roles played by those characters.

Experiment 1: Morphosyntactic and lexical verb information. In [9] subjects were presented with a scene showing, for example, a hare, a cabbage, a fox, and a distractor (see Fig. 1) together with either a spoken German SVO utterance (1) or with an OVS utterance (2):

(1) *Der Hase frisst gleich den Kohl.*
The hare$_{nom}$ eats shortly the cabbage$_{acc}$.
(2) *Den Hasen frisst gleich der Fuchs.*
The hare$_{acc}$ eats shortly the fox$_{nom}$.

The subject and object case-marking on the article of the first noun phrase together with verb meaning and world knowledge allowed anticipation of the correct post-

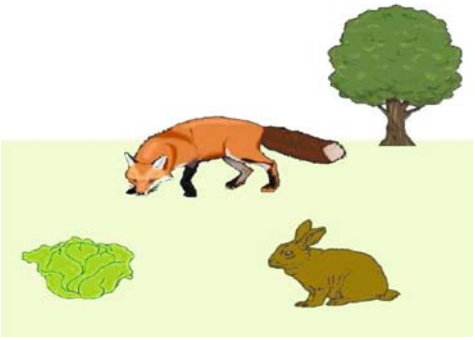


Fig. 1 Stereotypical Associations. People's world knowledge of stereotypical events allow them to anticipate upcoming arguments such as the fox in the figure based on an unfolding utterance *Den Hasen frisst gleich* ... ("The hare$_{acc}$ eats shortly ...")

verbal referent. People made anticipatory eye-movements to the cabbage after hearing "The hare$_{nom}$ eats ..." and to the fox after "The hare$_{acc}$ eats ...". Thus, when the utterance is unambiguous, and linguistic/world knowledge restricts the domain of potential referents in a scene, the comprehension system may predict mention of post-verbal referents.

Experiment 2: Verb type information. To further investigate the role of verb information, the authors used the same visual scenes in a follow-up study, but replaced the agent/patient verbs like *frisst* ("eats") with experiencer/theme verbs like *interessiert* ("interests"). The agent/experiencer and patient/theme roles from Experiment 1 were swapped. Given the same scene in Fig. 1 but the subject-first utterance (3) or object-first utterance (4), participants showed gaze fixations complementary to those in the first experiment, confirming that both syntactic case information and semantic verb information are used to predict subsequent referents.

(3) *Der Hase interessiert ganz besonders den Fuchs.*
The hare$_{nom}$ interests especially the fox$_{acc}$.
(4) *Den Hasen interessiert ganz besonders der Kohl.*
The hare$_{acc}$ interests especially the cabbage$_{nom}$.

2.2 Anticipation in Ambiguous Utterances

The second set of experiments investigated temporarily ambiguous German utterances. Findings showed that depicted events – just like world and linguistic knowledge in unambiguous utterances – can establish a scene character's role as agent or patient in the face of linguistic structural ambiguity (see Weber et al., conflicting constraints in Resource-Adaptive Language Comprehension of this volume, for further discussion of these experiments).

Experiment 3: Verb-mediated depicted events. For the sake of convenience, we will denote Experiment 3 by the acronym VMDE. In [11] comprehension of spoken utterances with local structural and thematic role ambiguity was investigated. An example of the German SVO/OVS ambiguity is the SVO utterance (5) versus the OVS utterance (6):

(5) *Die Princessin malt offensichtlich den Fechter.*
The princess$_{nom}$ paints obviously the fencer$_{acc}$.
(6) *Die Princessin wäscht offensichtlich der Pirat.*
The princess$_{acc}$ washes obviously the pirate$_{nom}$.

Together with the auditorily presented utterance, a scene was shown in which a princess both paints a fencer and is washed by a pirate (see Fig. 2). *Linguistic* disambiguation occurred on the second noun phrase (NP); in the absence of

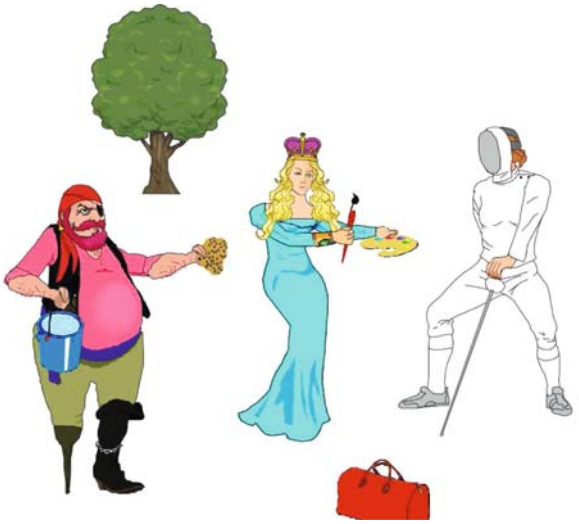


Fig. 2 VMDE: Verb-Mediated Depicted Events. When presented with an ambiguous sentence such as *Die Prinzessin wäscht gleich der* ... ("The princess$_{nom/acc}$ washes right away the ..."), participants use the utterance together with role information established by the relevant event in the scene to anticipate the *Pirat* ("pirate") as the most likely upcoming agent because the princess is the patient in the event that involves the action "washes"

stereotypical verb-argument relationships, disambiguation prior to the second NP was only possible through use of the depicted events and their associated depicted role relations. When the verb identified an action, the depicted role relations disambiguated towards either an SVO agent–patient (5) or OVS patient–agent role (6) relation, as indicated by anticipatory eye-movements to the patient (pirate) or agent (fencer), respectively, for (5) and (6). This gaze-pattern showed the rapid influence of verb-mediated depicted events on the assignment of a thematic role to a temporarily ambiguous utterance-initial noun phrase.

Experiment 4: Soft temporal adverb constraint. In [11] German verb-final active/passive constructions were also investigated. In both the active future tense (7) and the passive utterance (8), the initial subject noun phrase is role-ambiguous, and the auxiliary *wird* can have a passive or future interpretation.

(7) *Die Princessin wird sogleich den Pirat washen.*
The princess$_{nom}$ will right away wash the pirate$_{acc}$.
(8) *Die Princessin wird soeben von dem Fechter gemalt.*
The princess$_{acc}$ was just now painted by the fencer$_{nom}$.

To evoke early linguistic disambiguation, temporal adverbs biased the auxiliary *wird* toward either the future ("will") or passive ("is -ed") reading. Since the verb was utterance-final, the interplay of scene and linguistic cues (e.g., temporal adverbs)

were rather more subtle. When the listener heard a future-biased adverb such as *sogleich*, after the auxiliary *wird*, he interpreted the initial NP as an agent of a future construction, as evidenced by anticipatory eye-movements to the patient in the scene. Conversely, listeners interpreted the passive-biased construction with these roles exchanged.

Experiment 5: Relative priority of information type Again, for the sake of convenience, we will denote Experiment 5 by the acronum RPIT. In [10] a study built upon this research was presented that examined two issues. First, it replicated the finding that stored knowledge about likely role fillers of actions that were not depicted [9] and information from depicted (but non-stereotypical) events [11] each enable rapid thematic interpretation. An example scene showed a wizard spying on a pilot, to whom a detective is also serving food (see Fig. 3). For this experiment, item utterances had an unambiguous OVS order.

In two conditions (9 and 10) the first NP identified the central character in the scene, who is the patient of two events. When people heard the verb "jinx" in condition (9), *stereotypical knowledge* about jinxing identified the wizard as the only relevant agent, as indicated by a higher proportion of anticipatory eye movements to the stereotypical agent (wizard) than to the other agent. In contrast, when participants instead heard the verb "serves" in condition (10), the verb uniquely identified the detective as the only relevant agent via the scene depicting him in a food-serving event. Use of the depicted events was revealed by more inspections to the agent of

Fig. 3 RPIT: Relative Priority of Information Type. When presented with a sentence such as *Den Piloten bespitzelt gleich der* ... ("The pilot$_{acc}$ spies-on right away the ..."), participants could either look at the *Detektiv* ("detective") as the most likely upcoming agent based on its stereotypical association with the verb *bespitzelt* ("spies on"), or at the *Zauberer* ("wizard"), depicted as doing the spying. Empirical results show that people prefer the depicted event over stereotypical knowledge

the depicted event (detective) than to the other agent shortly after the verb. For future reference, we will call these conditions (9 and 10) the *No-Conflict* (NC) conditions, and distinguish them according to the available information source: *Stereo, No-Conflict* or *Scene, No-Conflict*.

(9) Stereo *Den Piloten verzaubert gleich der Zauberer*.
The pilot$_{acc}$ jinxes shortly the wizard$_{nom}$.
(10) Scene *Den Piloten verköstigt gleich der Detektiv*.
The pilot$_{acc}$ serves shortly the detective$_{nom}$.

Second, the study determined the *relative importance* of depicted events and verb-based thematic role knowledge. Participants heard utterances (11 and 12) where the verb identified both a depicted (wizard) or a stereotypical (detective) agent as relevant. When faced with this conflict, people preferentially relied upon the immediate event depiction over stereotypical knowledge for thematic interpretation. This was made evident by more looks to the wizard, the agent in the depicted event, than to the other, stereotypical agent of the spying action (the detective). Only when people heard the second NP was this interpretation revised as appropriate when the second NP mentioned the stereotypical role filler as the appropriate agent (11). This was reflected by a decrease of inspections to the wizard (the agent depicted as jinxing) and an increase of looks to the detective (the stereotypical agent) on NP2. Again, as we consider the modeling of these experiments, we will call these conditions (11 and 12) the *Conflict* (C) conditions, and distinguish them according to the information source that prevails on the final NP: *Stereo, Conflict* or *Scene, Conflict*:

(11) Stereo *Den Piloten bespitzelt gleich der Detektiv*.
The pilot$_{acc}$ spies-on shortly the detective$_{nom}$.
(12) Scene *Den Piloten bespitzelt gleich der Zauberer*.
The pilot$_{acc}$ spies-on shortly the wizard$_{nom}$.

2.3 Coordinated Interplay Account

In light of these findings of situated human language comprehension, particularly those of the VMDE and RPIT experiments, the *coordinated interplay account* (CIA) was proposed [10] as a theoretical framework reconciling the impact of various information sources on situated utterance understanding. The central insight of the CIA is that utterance-mediated attention in the scene not only "reflects" the process of incremental and anticipatory language comprehension, but that it also constitutes the mechanism by which the scene influences comprehension. The CIA stipulates that the interpretation of the unfolding utterance guides referential and anticipatory attention in the visual scene to establish reference to mentioned objects and events, and also to anticipate important scene regions or referents not yet mentioned.

As the utterance identifies relevant regions of the scene, attention shifts, and the visually attended information rapidly influences comprehension of the utterance. The CIA assumes tight coordination between when words are mentioned in an utterance, attention is shifted to relevant areas, and the attended region feeds back more specific information about the depicted event. This assumption is supported by the greater salience of attended scene information over linguistic and world knowledge. As it is likely that we all experience multiple sources of information that may be relevant to our understanding of language, we likely learn to prioritize them. In [10] it is argued that the relatively high priority of scene information may have a developmental explanation. Research in language acquisition suggests that the immediate environment plays an important role in a child's development as it learns to navigate the myriad possibly relevant stimuli to which it should respond first [16]. Evidence that concrete nouns and verbs are among the first concepts acquired by children, and that they relate to objects and perceptual events in their immediate environment suggest that both caregiver and child utilize a strategy that makes the search through possible referents more tractable once they reach adulthood. The grounding of these concrete words lays the foundation for the acquisition of more abstract concepts and relations that do not depend as greatly on the child's immediate environment, but is increasingly informed by their developing linguistic and world knowledge [7].

3 Connectionist Models

The two models we describe below are based on the Simple Recurrent Network (SRN; [6]) to produce a case-role interpretation of an input utterance as it is processed word by word. SRNs have become the mainstay in connectionist sequence processing because they are basically a standard feedforward network that incorporates dynamic context. We will use Fig. 4 below to exemplify sentence processing in SRNs; the SRN component of the network is labeled as such. Unlike symbolic or probabilistic models, SRNs process patterns (vectors) rather than symbolic representations. As mentioned in the Introduction, these types of networks derive their inspiration from an abstraction of how the brain works: massively interconnected simple processing units (often called neurons) that operate in parallel. These units are usually grouped into *layers* that themselves are an abstraction of the functional organization of the brain. These layers, in turn, may be partitioned into *assemblies* that are dedicated to specific functional tasks. Each unit in the network receives a weighted sum of the input units feeding into it, and outputs a value according to an activation function that generally is nonlinear in order to bound the output value in an interval such as [0,1]; indeed, most SRNs use the *logistic function*, $\sigma(x) = (1+e^{-x})^{-1}$, which is monotonically increasing and maps the entire real line to the interval [0,1]. The activation of a neuron is typically interpreted as its firing rate. SRNs are trained by providing an input sequence and a set of targets into which the network should transform the input sequence. The standard training algorithm is *backpropagation* and is an optimization technique that uses error signals derived

from the difference between the network's output and target to update the network weights to more closely approximate the targets on the next round of updates [15]. The weights between units could themselves grow without bound during training, but an input vector **x** transformed by the matrix of weights **W** to produce an output vector **y** that has been passed through the activation function σ ensures **y** remains bounded. In sum, for each pair of layers connected by a weight matrix, the output vector can be calculated simply as $\mathbf{y} = \sigma(\mathbf{Wx})$.

SRNs process sentences one word at a time, with each new input word represented in the *input layer* and interpreted in the context of the sentence processed so far – represented by the *context layer*, which is simply a copy of the *hidden layer* from the previous time step. The input layer and context layer are integrated and compressed into the hidden layer, and so the hidden layer represents the developing sentence. The *output layer* contains patterns the SRN has been trained to compute by providing targets for each output assembly (e.g., the *Verb* assembly in Fig. 4 that holds the output pattern for *wäscht* very closely approximates the targeted filler for the Verb role, *wäscht*).

The choice of a connectionist system was motivated by the cognitively plausible characteristics that these models automatically exhibit: They process sentences incrementally (Cognitive Characteristic 1), producing a developing interpretation of the unfolding sentence as output. Because these types of associationist models automatically develop correlations over the data they are trained on, they naturally develop expectations about the output even before a sentence is completely processed (Cognitive Characteristic 2). Connectionist models also performed better when they are trained to integrate multiple modalities [3, 4] (Cognitive Characteristic 3). Such a model can adapt its interpretation of the unfolding sentence to its semantic context [14] (Cognitive Characteristic 4). Moreover, these types of models develop behaviors such as difficulty with multiple center embeddings that often mimic those of people [2]. Finally, in [13] an architecture was presented that used an attentional mechanism to coordinate attention to objects and events in the scene with an incrementally processed utterance (Cognitive Characteristic 5). Due to space limitations, we can only highlight the two models' architectures and their input data, training and testing, and results. For more detailed information, the reader is directed to [12, 13].

3.1 Multimodal Integration Using Event Layers

The EVTNET model described in [12] was enhanced by a representation of scene information that was integrated into the model's processing (see Fig. 4). The encoding of the scene was complicated by the need to represent the depicted events, which involved actions, in addition to just the characters/objects themselves in the scene. Accordingly, the model had links from the characters to the hidden layer, links from the characters and depicted actions to *event* layers, and links from these event layers to the hidden layer of the SRN. Representations for the events were developed in

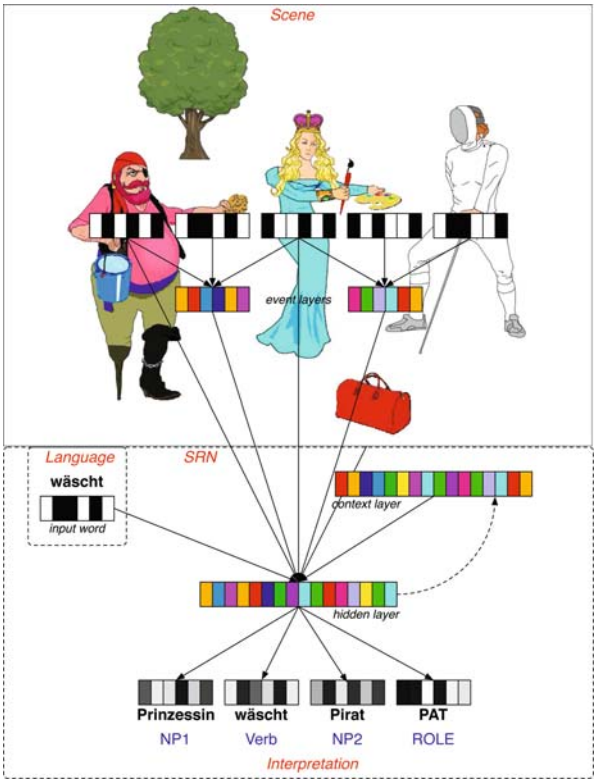


Fig. 4 Scene Integration via Compressed Event Layers. Representations of events in the scene are compressed into event representations in *event layers* through autoassociation of the events' constituents. The representations were then fed to EvtNet's hidden layer through shared weights. The representations of the characters themselves were also passed to the hidden layer to allow modeling experiments in which events were not explicitly depicted

the event layers by compressing the scene representations of the involved characters and depicted actions through shared weights corresponding to the action, the agent of the action, and the patient of the action. This event representation was kept simple to provide conceptual input to the hidden layer, divorced from the linguistic information the network was to learn from the utterance input. That is, who did what to whom was encoded for the events, when depicted; grammatical information came from the linguistic input.

3.1.1 Input Data, Training, and Testing

The network was trained to handle utterances based on Experiments 1–4 described in Sect. 2 involving both non-stereotypical and stereotypical events, as well as visual context when present or absent. We generated a training set of utterances based on the experimental materials while holding out the actual materials to be used for

testing. In order to accurately model the first two experiments involving selectional restrictions on verbs, two additional nouns were added for each verb that were superficially associated to the experimental materials. For example, in the utterance *Der Hase frisst gleich den Kohl*, the nouns *Hase1*, *Hase2*, *Kohl1*, and *Kohl2*[1] were used to develop training utterances so that the network could learn that *Hase*, *frisst*, and *Kohl* were correlated without ever encountering all three words in the same training utterance. The experiments involving non-stereotypicality did not pose this constraint, so training utterances were generated simply to avoid presenting experimental items. We made standard simplifications of lexical items by treating morphemes such as the infinitive marker *-en* and past participle *ge-* as separate words. All 326 words in the lexicon used in the first four experiments were given random binary representations (in which half of the units are on and the other half off) so that the network could not use features in the representations to solve its task. We tested the network by saving the epoch with the lowest training error, and computing performance results on the held-out test sets.

3.1.2 Results

As shown in Fig. 5, there are two points of primary interest in evaluating the performance of EVTNET: *anticipation* (ADV) of upcoming role fillers at the adverb and *comprehension* (NP2) of the complete utterance at the end of the utterance as observed experimentally. Both anticipation and comprehension are measured in terms of accuracy: For anticipation, we report the average percentage of predicted

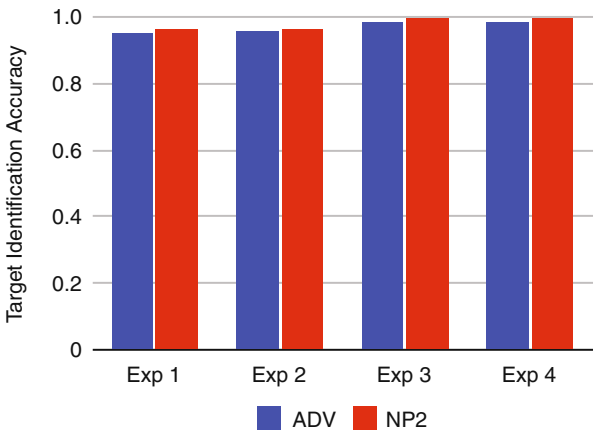


Fig. 5 Performance accuracy results for EVTNET on Experiments 1–4. The performance of EVTNET on each of the experiments 1-4 with respect to anticipation and comprehension all exceed 95%

[1] *Kohl1* and *Kohl2* could represent, for example, words such as "carrot" and "lettuce" in the lexicon that have the same distributional properties as *Kohl*, "cabbage".

upcoming targets in the test sets that the network correctly produces at the adverb in accordance with the empirical results. For comprehension, we similarly report the average percentage of correct interpretations for the test utterances. The model clearly demonstrates the qualitative behavior observed in all four experiments in that it is able to access the scene information and combine it with the incrementally presented utterance to anticipate forthcoming arguments.

For the two experiments using stereotypical information (Experiments 1 and 2), the network achieved just over 96% at the end of the utterance (NP2), and anticipation accuracy was approximately 95% at the adverb (ADV). Analysis shows that the network makes errors in token identification, confusing words that are within the selectionally restricted set of a given verb (e.g., *Kohl* and *Kohl2* with *frisst*). This confusion shows that the model has not quite mastered the stereotypical knowledge, particularly as it relates to the presence of the scene.

For the other two experiments using non-stereotypical characters and depicted events (Experiments 3 and 4), accuracy was 100% at the end of the utterance. More importantly, the model achieved over 98% early disambiguation on Experiment 3, where the utterances were simple, active SVO and OVS. Early disambiguation on Experiment 4 was somewhat harder because the adverb is the disambiguating point in the utterance as opposed to the verb in the other three experiments.

3.2 Multimodal Integration Using Attention

The EVTNET model demonstrated that connectionist systems could integrate a variety of information sources, including multimodal input from a scene. However, it fell well short of our goal to instantiate the CIA in which attention plays a crucial role. What was needed was a way of modeling attention directly, rather than inferring it from the network's output. Accordingly, we developed CIANET that features an explicit attentional mechanism, described below.

Based on the role that attention plays in the tight interaction of utterance and scene processing as described in the CIA, attention in CIANET was designed to be top-down utterance-driven: The input representations that the network processed served to indentify which of the two scene events was relevant for thematic interpretation. We used a *gating vector* (or gate) of the same size as the lexical assemblies (144 units) to implement an explicit attentional mechanism. The gate essentially transforms the architecture into a basic recurrent sigma-pi network [15], in which nodes may be multiplied as well as added together. The units of the gate are multiplied element-wise with the corresponding units in each of the three lexical representations comprising the agent, action, and patient of an event (see Fig. 6). To maintain the constraint that the more active one event is, the less active the other, each unit of the gate is subtracted from 1.0 to derive a vector complement that then modulates the other event's constituents. Crucially, the network is never explicitly taught to which event in the scene to attend. Rather, the gate is optimized to increase the contrast between the constituents of the two events based on error information

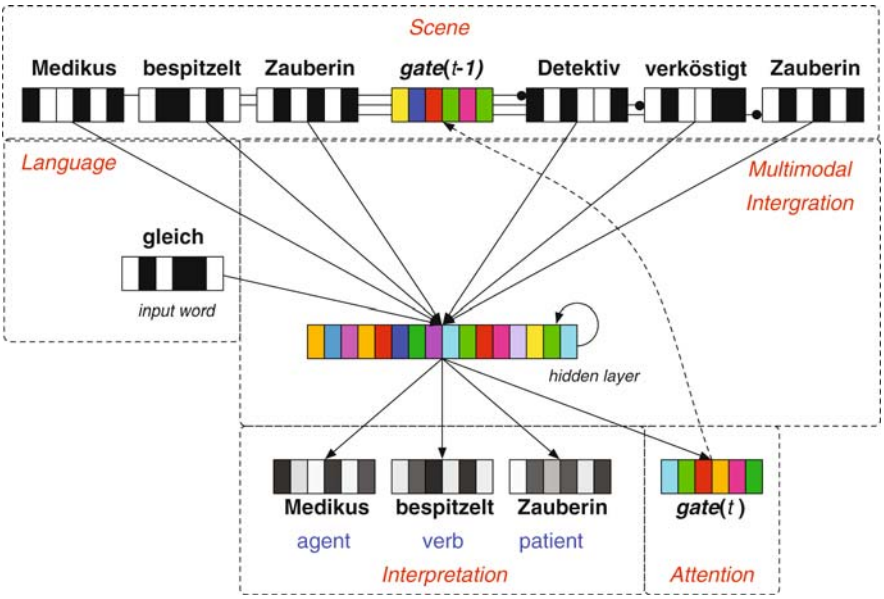


Fig. 6 Attention through Multiplicative Connections. CIANET learns to modulate each event in the scene through a gate that is multiplied element-wise with each constituent of an event. The *black circles* indicate that the gate's complement is multiplied element-wise against each constituent of the other event. At each step of processing, the gate is updated and the relative activations of the event's constituents are then passed to the hidden layer, where they are integrated with the current word and current utterance context

backpropagated recurrently during training from the multiplicative connections to the modulated constituent representations of each event. Consequently, the average activation of the gate's units directly correlates with greater activation of the attended event in a scene. Accordingly, attention is driven by correlations with the roles of arguments in the events and the linguistic aspects of the input utterance, such as case-marking and stereotypicality.

3.2.1 Input Data, Training, and Testing

We first describe how the lexical items, grammar, and scenes were developed to both train and test CIANET. A fundamental aim of CIANET was to investigate the scalability and generalizability of the model to more than a single experiment and to unseen conditions. To this end we selected the Verb-Mediated Depicted Events (VMDE) and the Relative Priority of Information Type (RPIT) experiments (Experiments 3 and 5, respectively) described in Sect. 2.

Testing: We constructed two test sets based on the linguistic and scene characteristics of the VMDE and RPIT experiments, and a third test set with SVO word order to complement the OVS structures of the RPIT experiment:

VMDE *Test Set:* The VMDE materials
RPIT *Test Set:* The RPIT materials with OVS word order
Balance Test Set: The RPIT materials with SVO word order

These test sets were held out from the training data along with corresponding validation sets. As is standard procedure in evaluating computational models, the lowest average error on the validation sets over the course of training the network is selected, and the performance statistics are computed on the held-out test sets to provide a measure of how well the system has generalized to unseen data.

Training: We trained CIANET using a generate and filter approach to utterances based on templates of these test sets. We also developed a common lexicon and grammar for both the VMDE and RPIT training and test sets to accommodate the experimental designs of the two experiments. A drawback of the earlier model was that the network could rely on lexical and grammatical distinctions between the experiments such as word order and scene composition to correlate utterances and scenes (e.g., learning that certain words always occurred as patients in both utterance and scene). The reason is due to the non-overlapping lexicons of each experiment. The approach of using a common lexicon forced the network to learn to coordinate linguistic, stereotypical, and non-linguistic information from the utterances and scenes. The lexicon consisted of 53 words (24 verbs and their associated, stereotypical agents, plus the three articles, adverb, and period). A quarter of the lexical items were feminine nouns that were only distinguished in the grammar by the article *die* because all lexical representations were given random binary values. The grammar had 26,208 utterances which could be paired with 91,570,176 scenes. In addition, we imposed a *No-Conflict Constraint* during training to filter out scenes with conflicting information sources to show that CIANET was not just fitting the data, but could also satisfy the third modeling goal of predicting the relative priority of depicted actions over stereotypical associations. In particular, the No-Conflict Constraint ensured that while the network is exposed to stimuli where either the scene or stereotypical information can predict role fillers, these two sources never occurred in the same scene (i.e., RPIT conditions 3 and 4 type stimuli were never seen during training, but were tested).

3.2.2 Results

We present the results by first measuring the overall performance of CIANET on anticipation of upcoming role fillers and comprehension of the utterance, as we did for the earlier model discussed above. We then examine how the attentional mechanism selects the event in the scene most relevant to the unfolding utterance. These results derive from ten runs of the network with validation sets and different random lexical encodings and seeds. In each run, half of the training utterances involve a stereotypical association between the agent and verb, and half do not. This 50% split, which we term the *stereotypicality ratio*, was based on the unbiased assumption that stereotypical associations and visual context occur in equal measure during language acquisition. This assumption is most likely not accurate,

and we will conclude the results and discussion with comparisons to simulations with different stereotypicality ratios biased in favor of the scene and stereotypicality, respectively.

Anticipation: The network performs with an overall accuracy above 99% on the situated utterances at the adverb on the VMDE Test Set 1 and RPIT Test set 2 for the No-Conflict conditions (e.g., *Die Zauberin bespitzelt gleich der Medikus* with the scene *[(Medikus bespitzelt Zauberin) (Detektiv verköstigt Zauberin)]* in Fig. 6). These are precisely the conditions CIANET has been exposed to during training. There is also a main effect in favor of the depicted agent (65%) versus the stereotypical agent (35%) in the Conflict conditions in the RPIT Test Set 2. This last result is statistically significant ($p < 0.0001$, $t = 13.0056$) and shows that the model both generalizes to novel input and correctly predicts the relative priority of depicted events over stereotypical associations. It does so at the adverb on the next step when the attentional mechanism selects the most relevant event by coordinating information from the utterance and scene in a strong temporal manner. The high numbers of accurate role fillers predicted shows that the network can anticipate role fillers based on depicted events or stereotypical information both for initially ambiguous utterances and for unambiguous utterances. Moreover, the ability of the network to adapt to available information (such as the presence of the scene) is reflected in the anticipation 93% of stereotypical agents when no scene is present.

Comprehension: CIANET also accurately maps the input utterance into its case-role representation at the end of utterance, achieving 100% of targeted roles in both the VMDE Test Set 1 and RPIT No-Conflict conditions in Test Set 2. The results of the model on the final thematic interpretation show that CIANET arrives at the correct interpretation for the VMDE and No-Conflict RPIT test sets. In particular, CIANET also captures the rapid revision that people perform in the stereotypical conflict condition when the NP2 forces them to revise the preference for the depicted agent towards the stereotypical agent. In the conflicting conditions in Test Set 2, the network correctly identifies the depicted agent in 92% of the cases, and the stereotypical agent in 81% of the cases. The errors the network makes derive from the fact that it is sometimes unable to override the agent it has anticipated at the adverb, a result that we will expound upon below when we examine the network's performance in more detail on an unfolding utterance.

CIANET *and Experimental Data.* We can compare CIANET's performance with the empirically observed behavior for both the VMDE and RPIT experiments both in terms of accuracy of anticipation of targeted objects in the scene (as measured at the ADV) and the network's accurate interpretation of the situated utterance at the end of sentence (measured at NP2). Figure 7 shows CIANET's performance compared with the renormalized experimental gaze proportions across each of the VMDE conditions. Similarly, Fig. 8 shows CIANET's performance compared with the renormalized experimental gaze proportions across each of the RPIT conditions. When comparing the experimental data with CIANET's performance, it is worth noting how similar the model's utterance-mediated anticipation of relevant role fillers is to human gaze patterns. Human gaze patterns and model prediction both exhibit correct anticipation of role fillers (agent for OVS and patient for SVO) in VMDE

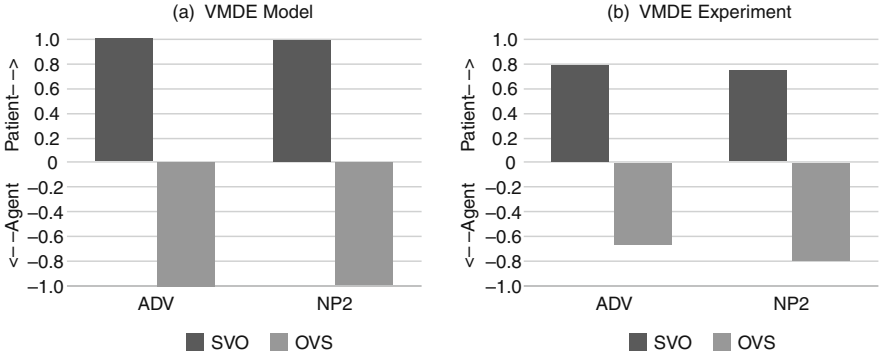


Fig. 7 Simulation vs. Empirical Behavior on the VMDE Test Set. CIANET's performance accuracy plotted on the *y*-axis (**a**) compares qualitatively with the empirical behavior observed in the VMDE experiment (**b**) both with respect to anticipation of upcoming role fillers on the ADV and comprehension once NP2 has been processed

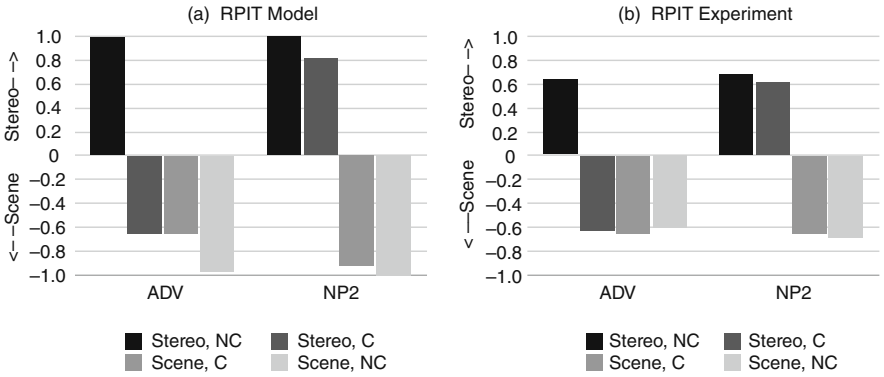


Fig. 8 Simulation vs. Empirical Behavior on the RPIT Test Set. Similarly, CIANET (**a**) shows qualitative agreement on the RPIT conditions compared to the RPIT experiment (**b**). In the No-Conflict conditions, the depicted agent is preferred because it is the only source of information, whereas the stereotypical agent is preferred when the utterance contains an associated verb. In the Conflict conditions, the depicted agent is preferred because the verb in the utterance mediates the corresponding depicted action of which the agent is a constituent. Accuracy is plotted on the *y*-axis

as well as in the No-Conflict conditions of RPIT (depicted agent in depicted target condition and sterotypical agent in the No-Conflict stereotypical target condition). Even more interesting is the fact that – just as people do – the model predicts the preference for the depicted agent in the Conflict RPIT conditions on which the network has not been trained. The fit of the model to the empirical data was assessed using the non-parametric rank-order Spearman's ρ with $\rho = 0.9209$ (two-tailed, $p < 0.00002, t = 7.47$).

To give a more fine-grained picture of the network's behavior, we now examine the processing of an unfolding utterance from Test Set 1 (VMDE) and Test Set 2 (RPIT) in more detail.

VMDE: Fig. 9 plots the average activation of the attention vector together with error bars denoting the average variance across the ten simulations on Test Set 1. As described in Sect. 3.2, CIANET's attentional mechanism modulates the activations of the events in the scene. Intuitively, if the elements of the attention (gating) vector are all close to one or all close to zero, then one event in the scene will be highly activated at the expense of the other event. Figure 9 shows that this intuition is very close to how CIANET actually operates. The x-axis shows an utterance template for VMDE experimental materials, and the y-axis relates the average activation of the attention vector to whether the upcoming role filler is an agent for OVS or a patient for SVO word order. For the first NP *die Noun1*, the mean is essentially 0.5 with a lot of variance at the noun because the gating vector has not accumulated enough information to select either event as most relevant to the input utterance. However, once the *Verb* is read in, the mean of the gating vector immediately reflects the relevant event based on identification of the utterance verb with the corresponding action in the scene, and the substantial fall in variance indicates that the vector is operating essentially as a single multiplicative scalar. It is important to note that the change in activation of the attention mechanism occurs at the verb, but its effect on the event representations in the scene can only occur on the next step once the attention vector is copied to modulate those events in the scene.

Figure 10 demonstrates how attention shifts over the course of processing an utterance on the two VMDE conditions in Test Set 1. The plot shows the difference between the normalized Euclidean distances of CIANET's output to the targeted role

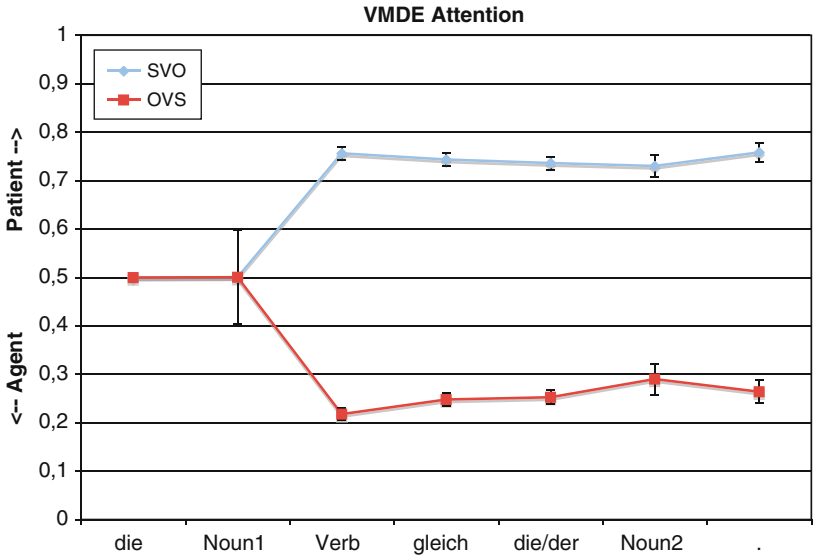


Fig. 9 Attention on VMDE Test Set. The average mean and variance of the gating vector is plotted for the two VMDE conditions. Once the verb is read in, the network selects the most relevant event to the utterance processed up to that point. Gate activation is plotted on the y-axis

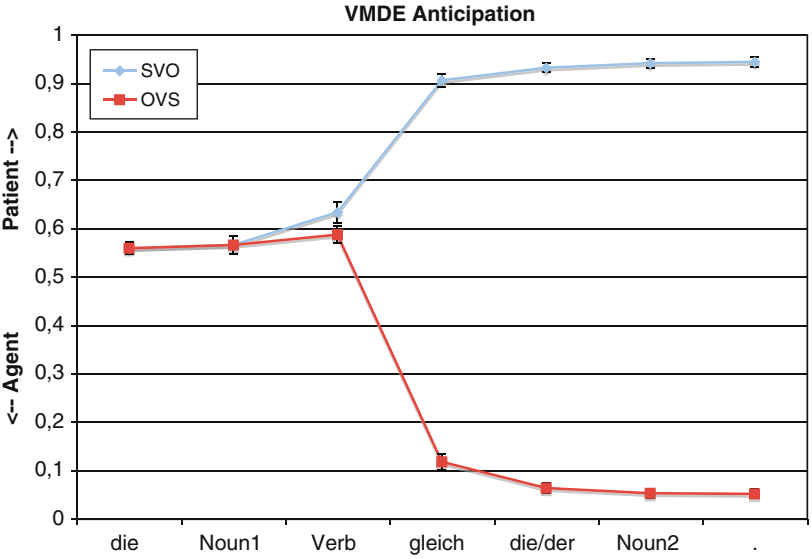


Fig. 10 Anticipation on VMDE Test Set. The influence of the gate is shown for the two VMDE conditions. Once the gate has identified the relevant action in the scene, the anticipation of the upcoming role filler (Agent vs. Patient) is reflected in the average difference of the normalized distances between the CIANET's output and the eligible fillers in the scene, as plotted on the *y*-axis

fillers in the scene. These values were collected for all test utterances and averaged, and then translated to lie between 0 and 1 to make comparison with Fig. 9 easier. The results show a slight bias toward SVO utterances because the network initially predicts the common (feminine) character in the scene as both agent and patient in the output interpretation, albeit very weakly. This filling of two roles with the same argument is typical of SRNs trained to output a case-role interpretation, and may be interpreted as the network's maintenance of both possibilities in parallel until disambiguating information is encountered. Once it has read in the verb, the attention mechanism then activates the relevant event on the next step (the adverb *gleich*, at which point the thematic role of the first NP is clear). The results are biased toward SVO and not offset by OVS on the first NP because scenes in which a common agent is acting on two distinct patients never occur as scenes during training or testing in the simulations.

RPIT**:** Fig. 11 shows the plots of the mean and variance of the units in the gating vector over the four RPIT conditions in Test Set 2. The plots have been arranged so that a mean greater than 0.5 shows a preference for the event with the stereotypical agent in the scene, whereas a mean less than 0.5 indicates a preference for the event with the verb-mediated depicted agent. The close parallels between Figs. 11 and 12 below suggests that it is indeed the attentional mechanism that is driving the dynamics of the model. The behavior of the attention vector up to the verb is the same as in Fig. 9 for the VMDE test set. But the dynamics differ after the verb according to condition. Figure 11 shows that only in the "Stereo, No-Conflict" condition does

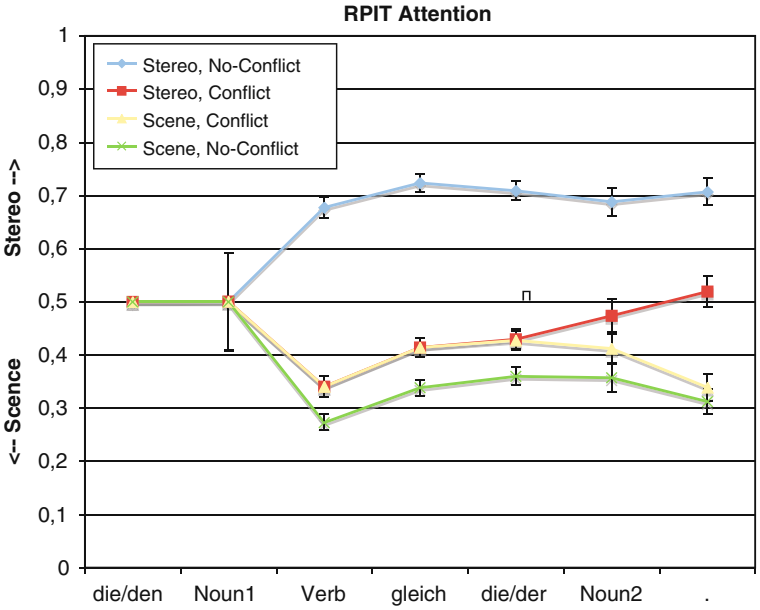


Fig. 11 Attention on RPIT Test Set. As in Fig. 9, the average mean and variance of the gating vector is plotted for each of the four conditions in the RPIT Test Set. Activation is plotted on the *y*-axis

attention select the event with the stereotypical agent as most relevant. In all the other conditions, the event with an action matching the processed verb is selected, with the result that the agent of that depicted action is activated more than the other event in the scene. This agent preference is then reflected on the next step at the adverb *gleich*, as will be shown in Fig. 12 below. It is also important to note that the attentional mechanism does not become fixated on an event, but rather that the mean slowly decays back toward 0.5 as the rest of the utterance is processed. The reason is that processing still has to accommodate the final noun phrase, where the actual input agent may differ (in the "Stereo, Conflict" condition) with the agent of the event the network had attended to, and the network has learned to accommodate this possibility. Once the final agent is read, attention rebounds toward that agent on the end-of-utterance period. As in Fig. 9, the influence of the attention vector takes effect on the next step to enable the anticipation of the upcoming role agent.

Figure 12 gives a clearer view of how attention shifts over the course of processing an utterance for all four conditions of Test Set 2. The plots show the average difference between the normalized Euclidean distances of the network's agent output and the stereotypical and depicted (scene) agents. As in Fig. 11, the values have been transformed so that a value greater than 0.5 indicates that the output is closer to the stereotypical agent, and a value less than 0.5, to the depicted agent. The dashed line in the middle of the graph is the average of the conditions on which the network was trained (the No-Conflict conditions), and shows that there

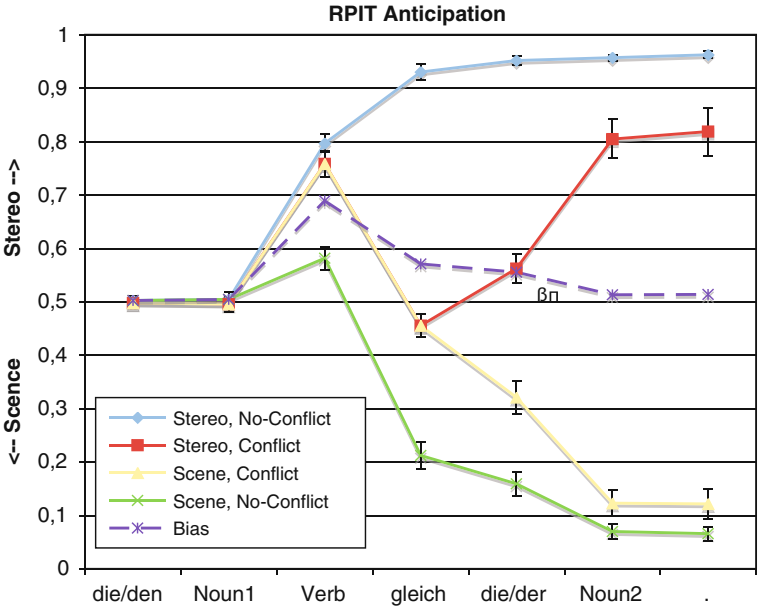


Fig. 12 Anticipation on RPIT Test Set: the coordinated interplay of information sources. The varying preference for the stereotypical (Stereo) versus depicted (Scene) agent averaged for each of the four conditions over the test set clearly shows the model's ability to adapt to information as it processes a sentence incrementally. The difference in normalized Euclidean distance is plotted on the *y*-axis

is a distinctive stereotypical bias after the first NP due to the influence of the utility of stereotypical information even when no scene is present. The average value at the adverb is 0.4552, which when measured against a hypothetical mean of 0.5 is statistically significant ($p < 0.0001$, $t = 9.4465$), and when measured against the inherent stereotypical bias of 0.571 as indicated by the dashed line at the adverb, is also statistically significant ($p < 0.0001$, $t = 22.8995$). Thus, the two Conflict conditions demonstrate the network's preference for the depicted agent at the adverb, overcoming the stereotypical bias at the verb. We will step through the plots in Fig. 12 as an utterance unfolds, highlighting the pertinent states of each condition that demonstrate the influence of the basic attentional mechanism:

1. *die/den Noun1:* As it processes the first NP, CIANET initially shows no preference for either event agent because the patient appears in both events and is not strongly correlated with either agent.
2. *Verb:* A preference for the stereotypical agent over all conditions is evident when the model has just processed the input verb, but not yet shifted attention to the most relevant event. This behavior is a prediction of CIANET amenable to experimental verification.

- This initial preference for the stereotypical agent makes sense for the conflicting conditions "Stereo, Conflict" and "Scene, Conflict" because a stereotypical agent does appear in the scene, and, in the case of the "Stereo, Conflict" and "Stereo, No-Conflict", is strengthened by the attentional mechanism predicting the event in which the stereotypical agent occurs is relevant.
- For the "Scene, No-Conflict" condition, however, it may reflect a negative correlation between the input verb and verb-mediated depicted agent that has developed as an artifact of the limited number of 24 verbs and their associated nouns used in the study.

3. *Adverb:* The bias towards the stereotypical agent is overridden on the next step (the adverb) once the network has shifted attention to the most relevant event.

   - For the "Scene, No-Conflict" condition, in which only case-marking on the first NP and thematic role information from the processed verb combine with information from the depicted event, there is a strong preference for the depicted agent. Despite the presence of an ambiguous feminine noun in the first NP in some of the test utterances, the depicted information and processed verb is enough for the network to anticipate the correct upcoming agent.
   - For the "Stereo, No-Conflict" condition, there is in contrast a strong anticipation of the stereotypical agent since the processed input verb has no corresponding depicted action.
   - The interaction between the unfolding input utterance and the scene is evident in the network's varying anticipation of the upcoming role filler in the two (untrained) conflicting conditions, which are identical up to the adverb. The network does show a clear anticipation of the depicted over the stereotypical agent at the adverb *gleich*.

4. *den/die Noun2:* Finally, the zigzag form of the "Stereo, Conflict" curve in Fig. 12 attests to the ability of CIANET to rapidly adapt to information as it becomes available: at the *Verb*, stereotypicality is the most informative source, and that is integrated with information from the scene on the next step to shift attention to the relevant event supporting anticipation of the verb-based depicted agent, but finally overridden on the final NP, which contains the stereotypical agent.

In both of the VMDE and RPIT experiments, CIANET is able to identify the most relevant event in the scene and activate it more highly over the irrelevant event. It is able to do so despite the variety of information sources it must use, such as stereotypicality, case-marking, argument structure, and plausibility. This ability strongly suggests that the system is able to navigate these various information sources and prioritize them to produce the correct interpretation based on its experience during training. The system assigns the correct thematic roles to the words in the utterance not only at the end of the utterance, as would be expected, but also demonstrates anticipation at the adverb by assigning the thematic role observed in the experimental studies. Recall from Sect. 2 that people revise their interpretation once they have heard the entire utterance. CIANET mimics this behavior as well.

The role of experience in CIANET*:* Finally, we present results that highlight the role of experience on the priority of the stereotypical knowledge and the immediate visual context. Table 1 shows comparisons of CIANET's anticipatory performance over each RPIT condition for stereotypicality ratios of 25, 50, and 75% respectively, as measured at the adverb. To illustrate the concept of stereotypicality ratio: For a stereotypicality ratio of 25%, CIANET is exposed to a stereotypical agent in 25% of the training set utterances. Performance on the No-Conflict conditions are comparable for each ratio, although there is a noticeable preference for the stereotypical agent vs. depicted agent (100% vs. 88%) for the stereotypicality ratio of 75%. The contrast is even more evident on the Conflict conditions, but it is in keeping with what would be expected. At the stereotypicality ratio of 25%, information from the scene occurs three times more frequently than does stereotypical associations, and its influence is evident in the 85% accuracy for anticipation of upcoming agents. At this ratio of 25%, the stereotypical agent occurs with its corresponding verb roughly 6.6 times as often as with any other agent during training, so the association is not negligible. The results for the 50% ratio have already been presented, where, despite the fact that associations between a verb and its stereotypical agent are 23 times more frequent than nonstereotypical associations (those distributed across the other 23 non-stereotypical agents), the network still demonstrates a noticeable preference for the depicted agent in the scene. At a ratio of 75%, stereotypical ratios are 67 times more frequent than information from the depicted agent in the scene, and that influence is reflected in the network's correspondingly strong preference of the stereotypical agent (82%) over the depicted agent (18%). Bear in mind that only the relative frequency of stereotypical vs. depicted agents were manipulated; the scene itself was still presented half of the time. For the purposes of completeness, the network exceeded 97% accuracy on comprehension at the end of utterance. When the scene was not present, the influence of the three ratios were also evident: at 25%, only 76% of stereotypical agents were correctly anticipated in OVS utterances; at 50%, 93% were anticipated, and at 75%, over 99% were correctly anticipated.

Table 1 Influence of experience with visual context vs. stereotypical knowledge on utterance interpretation

	Stereotypicality ratio											
	25%				50%				75%			
RPIT	No-conflict		Conflict		No-conflict		Conflict		No-conflict		Conflict	
	Dep.	Ster.	Dep.	Ster.	Dep.	Ster.	Dep.	Ster.	Dep.	Ster.	Dep.	Ster.
Stereo	0.8%	99.2%	74.7%	25.3%	0.4%	99.6%	65.3%	34.7%	11.6%	88.4%	27.6%	72.4%
Scene	99.0%	1.0%	74.7%	25.3%	97.2%	2.8%	65.3%	34.7%	100%	0.0%	27.6%	72.4%

4 Conclusion

We have described the evolution of a connectionist model from an architecture using event layers, EvtNet, to a more parsimonious recurrent sigma-pi neural network architecture, CIANET, motivated by the coordinated interplay account (CIA) of situated utterance comprehension. The CIA led, in turn, to the formulation of the *Modeling Goals* that a model featuring interlocking language–scene coordination should possess. CIANET exhibits the five *Cognitive Characteristics* (Goal 1) identified in Sect. 1: The model operates (1) incrementally, processing each word in the utterance in context. It also (2) accurately anticipates upcoming role fillers based on either stereotypical knowledge or information from the scene by seamlessly (3) integrating these information sources together with linguistic information from the input utterance. Moreover, the model is (4) adaptive, able to perform correctly when there is no scene, and, in general, uses information as it becomes available, such as verb-based stereotypical thematic knowledge. Finally, the manner in which the events are selected by the attention mechanism based on the utterance verb and manifested at the adverb demonstrates the (5) coordinated interplay between utterance and attention to the scene. This behavior can be seen as directly instantiating the CIA: the utterance directs the attentional mechanism of the network to select the most relevant event in the scene to the utterance, which then directly influences the network's full interpretation, as revealed by what it anticipates and when.

The cognitive behavior of CIANET results from the dynamic activation of relevant events in the scene through an *explicit* attentional mechanism as the utterance is processed (Goal 2). Either the first NP (in the case of stereotypicality) or the main verb (for depicted actions) may influence the selection of the relevant event, with the resulting influence of attention delayed one time step. The attentional mechanism is implemented by a gate that modulates the two events fed into the SRN through shared weights. The gate then selects the appropriate event in the scene by contrasting the relevant event's constituents and those of the irrelevant event through implicit inhibition.

The primary *empirical* result of the simulations presented in this chapter is that the network correctly learns to resolve conflicting information sources in favor of the immediate scene over stereotypical knowledge, despite only being trained on non-conflicting utterances (Goal 3). This result comes from the static presence of the scene that facilitates the correlation of the verb in the utterance with a depicted action in one of the events in the scene, whereas the stereotypical information only comes into play once the verb or its stereotypical agent is processed. Because the network must learn to identify and attend to the relevant event in the scene, its relative influence becomes amplified with training. Finally, the manner in which CIANET learns to attend to relevant events in the scene highlights the *developmental* aspect of situated language understanding (Goal 4). The model learns to attend to relevant objects and actions in the scene, when present, and uses correlations among them to anticipate upcoming role fillers. The use of the attentional mecha-

nism allows CIANET to be directly compared with human behavior as observed in psycholinguistics experiments. The resulting behavior accords well with evidence that attention – at least at the cellular level – also works by increasing the distinction among stimuli [19].

In summary, the final architecture, CIANET, meets each of the Modeling Goals set forth in Sect. 1. It exhibits the Cognitive Characteristics of incrementality, anticipation, adaptiveness, integration, and temporal coordination. The model also uses an explicit attentional mechanism to increase the salience of the event in a scene most relevant to an utterance being processed. Furthermore, it exhibits the preference for the immediate visual context as observed in human subjects. Lastly, this preference is dependent on the relative frequencies of stereotypical and depicted events that CIANET was exposed to during training, suggesting that developmental factors play a role in the experimental findings. The experiments modeled here posed a number of challenges for CIANET: It had to be able to correctly anticipate upcoming role fillers based on depicted events in both initially structurally ambiguous utterances and utterances in which grammatical function was clear from case marking on determiners. It also had to prioritize linguistic, non-linguistic, and stereotypical knowledge in a manner consistent with that observed in people. Lastly, the model had to be able to potentially override an anticipated role filler at the end of the utterance when the final NP differed. Thus, CIANET represents a significant step toward modeling situated language understanding in which a variety of linguistic and non-linguistic information sources interact to produce an interpretation of the utterance as it is processed.

References

1. Altmann, G.T.M., Kamide, Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3):247–264 (1999).
2. Christiansen, M.H., Chater, N. Toward a connectionist model of recursion in human linguistic performance. Cognitive Science, 23(2):157–205 (1999).
3. Christiansen, M.H., Conway, C.M., Curtin, S. Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In J.W. Minett, W.S.Y. Wang (Eds.), Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics. Hong Kong: City University of Hong Kong Press (2005).
4. Conway, C.M., Christiansen, M.H. Modality constrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31:24–39 (2005).
5. Cooper, R. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology, 6:84–107 (1974).
6. Elman, J.L. Finding structure in time. Cognitive Science, 14(2):179–211 (1990).
7. Gillette, J., Gleitman, H., Gleitman, L., Lederer, A. Human simulations of vocabulary learning. Cognition, 73:135–176 (1999).
8. Kamide, Y., Altmann, G.T., Haywood, S. Prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, 49:133–156 (2003).
9. Kamide, Y., Scheepers, C., Altmann, G.T. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. Journal of Psycholinguistic Research, 32(1):37–55 (2003).

10. Knoeferle, P., Crocker, M.W. The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye-tracking. Cognitive Science, 30(3):481–529 (2006).
11. Knoeferle, P., Crocker, M.W., Scheepers, C., Pickering, M.J. The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. Cognition, 95:95–127 (2005).
12. Mayberry, M.R., Crocker, M.W., Knoeferle, P. A connectionist model of sentence comprehension in visual worlds. In Proceedings of the 27th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ (2005).
13. Mayberry, M.R., Crocker, M.W., Knoeferle, P. A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In Proceedings of the 28th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ (2006).
14. Mayberry, M.R., Miikkulainen, R. Lexical disambiguation based on distributed representations of context frequency. In Proceedings of the 16th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ (1994).
15. Rumelhart, D.E., Hinton, G.E., Williams, R.J. Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations, (pp. 318–362). Cambridge, MA: MIT Press (1986).
16. Snow, C.E. Mothers' speech research: From input to interaction. In C. Snow, C. Ferguson (Eds.), Talking to Children: Language Input and Acquisition. Cambridge, MA: Cambridge University Press (1977).
17. Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M., Sedivy, J.C. Eye-movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. Cognitive Psychology, 45:447–481 (2002).
18. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C. Integration of visual and linguistic information in spoken language comprehension. Science, 268:1632–1634 (1995).
19. Taylor, J., Hartley, M., Taylor, N. Attention as sigma-pi controlled ACh-based feedback. In Proceedings of the International Joint Conference of Neural Networks. Elsevier Science Ltd., Oxford, UK (2005).

Part II
Resource-Adaptive Processes in Human–Machine Interaction

Assessment of a User's Time Pressure and Cognitive Load on the Basis of Features of Speech

Anthony Jameson, Juergen Kiefer, Christian Müller, Barbara Großmann-Hutter, Frank Wittig, and Ralf Rummer

1 Introduction

The project READY (1996–2004) approached the topic of resource-adaptive cognitive processes from a different angle than most of the other projects represented in this volume: The resources in question were the cognitive resources of computer users; the adaptation was done by the system that they were using.

The type of system focused on in the research was mobile conversational systems, for reasons that will become clear below. The resource limitations of interest concerned the user's available time and working memory.

Since it would be impractical to discuss all of the lines of research in the project within a single chapter, this chapter will focus on one issue that was addressed in a number of studies over a period of several years, including one study whose results have not been published previously: the issue of how a system can estimate the time pressure and cognitive load of its user, in particular on the basis of evidence in the user's behavior with the system, such as their speech.

In passing, we will also mention some of the related work in the READY project, as well as other related research. Other aspects of the research in READY, especially concerning the use of probabilistic methods for user modeling, are discussed in the chapter by Wittig in this volume.

1.1 Reasons for Variation in Cognitive Load and Time Pressure

One salient issue in the design of mobile conversational interfaces is the role of situationally determined *resource limitations* of the user – specifically, time pressure and cognitive load.

A. Jameson (✉)
DFKI GmbH, Saarbrücken, Germany; Fondazione Bruno Kessler – Istituto per Ricerca Scientifica e Tecnologica (FBK-irst), Trento, Italy
e-mail: jameson@dfki.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_9,

Compared with the users of stationary interactive systems, mobile users are more likely to be experiencing environmentally induced cognitive load. The user $\mathcal{U}$'s attention to the environment may be due simply to distracting stimuli in the environment (as when $\mathcal{U}$ is being driven in a taxicab while using the system $\mathcal{S}$)[1]; but often $\mathcal{U}$ will be attending actively to the environment while performing actions in it (e.g., handling objects or navigating through the environment). The tendency of users to attend to their environment and to multitask may be even greater with conversational mobile systems than with those that do not use speech as a communication channel, because of the largely eyes-free and hands-free character of speech.

Although users of stationary systems can of course also experience time pressure, especially acute time pressure can arise when a conversational interface is used during interaction with other persons or the environment. For example, a driver may want to complete a task while waiting at a stoplight; or a user may be interacting with another person who herself has little time available.

Research on how designers of technical devices can take situationally determined resource limitations into account has a long tradition in the field of engineering psychology (see, e.g., [1]). In the airplane cockpit, the automobile, or the nuclear power plant, the importance of factors like mental load and time pressure is too obvious to be overlooked. The research of this sort that seems most directly relevant to mobile conversational systems is research on in-car systems for drivers (see, e.g., [2, 3]). The advent of conversational systems for drivers has been motivated largely by the perceived fundamental compatibility of speech with the task of driving (see, e.g., [4]).

With other types of mobile conversational interface, research on the role of user resource limitations is still in a relatively early stage. But it would be inappropriate to neglect them. Consider, for concreteness, the example of a conversational system that serves as an assistant to a traveler in a large airport, answering questions and providing guidance. Figure 1 illustrates how quite different system behaviors may be appropriate given different user resource limitations.

1.2 Why Automatic Adaptation?

There are, of course, straightforward ways of ensuring that a system shows appropriate behaviors in cases like this. First, the user could be allowed to specify explicitly what type of system response they prefer – for example, by including in the spoken query the request for a response that contains only the minimally necessary information. But especially when the user's resources are limited, such explicit specification may require too much mental effort and/or time. Second, the designers of the system can try to ensure that its basic design makes it highly usable even given severe

[1] To simplify exposition, we will use the symbols $\mathcal{S}$ and $\mathcal{U}$ to denote a system and its user, respectively.

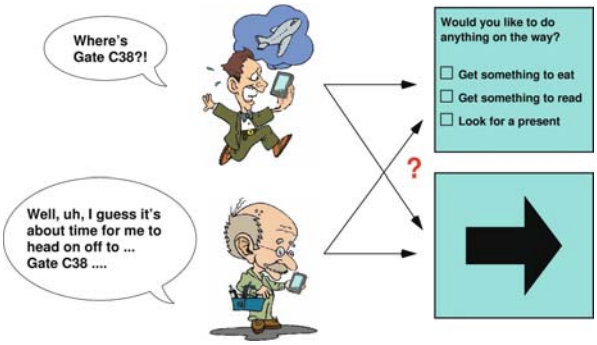


Fig. 1 Example of how a user's current resource limitations can call for different system responses. (Each of the two screens shown is a possible system response to the user's input utterance.)

resource limitations – for example, by providing only simple displays such as the lower one in Fig. 1. But a design that is well suited for one particular combination of resource limitations may not be well suited to a different combination, or to a situation in which there are no significant limitations. For example, the minimalistic output on the lower screen in Fig. 1 is unlikely to be optimal for the second user. And even the user experiencing time pressure might prefer a different type of display if he is not also experiencing cognitive load.

One possible approach to this dilemma is to give the system some capability to recognize the user's resource limitations automatically and to adapt to them with some degree of autonomy. In the next section, we will give some further examples of how this type of adaptation can be appropriate. Section 3 will then consider the first question that this approach raises – How can a system automatically recognize resource limitations? – giving an overview of possible methods. Against this general background, the remaining major sections of the paper will present specific empirical results and analyses concerning the role of the user's speech as a source of evidence on which adaptation to resource limitations can be based.

2 Possible Forms of Adaptation

Let us suppose in this section that a mobile conversational interface $\mathcal{S}$ is capable of making some reasonably accurate estimate of the user $\mathcal{U}$'s resource limitations at a given moment. How might $\mathcal{S}$ make use of this assessment to generate more appropriate system behavior? If there are no plausible answers to this question, there is little point in investigating techniques for assessing resource limitations.

2.1 Interruption of Communication

Perhaps the simplest form of adaptation is for $\mathcal{S}$ simply to stop communicating with $\mathcal{U}$ when $\mathcal{S}$ perceives resource limitations. For example, [3] describes a prototype

conversational in-car navigation system that interrupts its speech output whenever the driver applies the brakes. The goal is that in critical traffic situations, $\mathcal{U}$ should be able to devote their full attention to the driving task. In effect, the depression of the brake pedal is being interpreted as an indicator of high cognitive load.

2.2 Timing and Form of Notifications

Some conversational systems spontaneously present notifications to users. For example, the wearable NOMADIC RADIO [5] transmits audio messages (such as voice mail) to the user in a context-sensitive fashion. Although NOMADIC RADIO does not explicitly model $\mathcal{U}$'s cognitive load or time pressure, it does take into account related factors, such as whether $\mathcal{U}$ is currently interacting with $\mathcal{S}$ and whether $\mathcal{U}$ is in a meeting. In addition to postponing notifications, the system can choose from several forms of notification that have different degrees of obtrusiveness.

Other notification systems that assess the user's context have been presented by Horvitz and colleagues (see, e.g., [6, 7]). These systems make use of decision-theoretic methods to weigh the benefits of a notification against the costs (e.g., distraction). Here again, cognitive load and time pressure are not modeled explicitly.

2.3 Dialog Strategy

Many conversational systems are capable of switching between different dialog styles depending on the current state of the interaction. For example, [8] describes TOOT, a prototype spoken dialog system for retrieving online train schedules. TOOT sometimes applies a highly conservative dialog strategy in which each piece of required information (e.g., destination, place of departure, and time of departure) is elicited from the user through a focused question and then confirmed through a yes–no question. With less conservative strategies, $\mathcal{S}$ asks more open questions that allow $\mathcal{U}$ to specify two or more pieces of information at a time (e.g., "How may I help you?"). $\mathcal{S}$ decides which strategy to use on the basis of features of the current dialog, such as the system's confidence in the success of its own speech recognition. The main motivation here is to allow users whose speech can be recognized relatively well to proceed through the dialog quickly, while still accommodating users whose speech is problematic. But analogous changes in dialog strategy could be based on assessments of cognitive load and/or time pressure: The more conservative strategies may be especially appropriate for users who are currently distracted by the environment or by another task, whereas they may be especially frustrating for users under time pressure.

Such hypotheses about the suitability of particular dialog styles for particular configurations of resource limitations of course require a theoretical and empirical foundation. An effort along these lines was made in a different line of research in the

READY project [9]: In an experimental setting, each of 24 subjects used a mouse to carry out spoken instructions regarding a graphical control panel (e.g., "Set *X* to 3, set *M* to 1, set *V* to 4"). In half of the trials, the instructions for a given panel were *bundled*, as in the example just given; in the other half of the trials, they were presented *stepwise*: After each single instruction (e.g., "Set *X* to 3"), the system waited until the user had completed the instruction and clicked on a confirmation button; then the system presented the next individual instruction. An orthogonal manipulation induced cognitive load in half of the trials through a secondary task that required subjects to attend to color changes in one part of the screen.

When instructions were presented bundled, subjects often made errors when a sequence comprised 3 or 4 instructions and when they were distracted by a secondary task. By contrast, the stepwise presentation of instructions was shown to be a slow but safe strategy, like the conservative dialog strategies discussed above: Subjects made very few errors even in the most difficult conditions. Given the assumption that users attach some value to both rapid task completion and the avoidance of errors, it can be shown that stepwise presentation is on the whole relatively suitable when $\mathcal{U}$ is experiencing cognitive load, but that the system's choice between the two modes should also be based on the length of the instruction sequence and the relative importance of execution speed and error avoidance. Although it was conducted in an artificial environment, this study empirically confirms the intuition that different dialog strategies can be suitable under different configurations of resource limitations.

2.4 Other Forms of Adaptation to Resource Limitations

Several other ways in which a conversational interface might adapt to resource limitations should be mentioned briefly for completeness, although they so far have been instantiated less clearly than the possibilities discussed above.

On the basis of perceived high cognitive load, a system might change its behavior as follows:

- Present a smaller amount of optional information that is not strictly required for the performance of $\mathcal{U}$'s system-related task.
  For example, the airport assistant introduced above might stick to basic navigation instructions while guiding $\mathcal{U}$ from one location to another, leaving out information about airport facilities passed along the way.
- Present information in a style that is optimized for easy understanding, at the expense of other criteria (such as elegance or conciseness).
  Some stylistic features (e.g., simplicity and explicitness) are commonly recommended for texts that are typically read or heard by users who cannot be expected to be paying full attention, such as error messages and help texts (see, e.g., [1], Chap. 6). The novel idea in an adaptive system is that the degree to which such elements should be included should depend on the perceived level of cognitive load, because of the trade-offs with other criteria.

- Adapt the interface in such a way as to prevent errors that are typical of high cognitive load.
  A number of categories of *expert slip* are discussed by Norman [10], along with design remedies. Each such remedy (e.g., making objects more visually distinctive; asking for confirmation) tends to have some drawbacks. Since expert slips are especially likely when $\mathcal{U}$ is environmentally distracted, some remedies may become worthwhile under high cognitive load even if their drawbacks outweigh their advantages given low cognitive load.

Analogous suitable responses to time pressure might include the following:

- Present concrete instructions that describe specific actions, as opposed to encouraging $\mathcal{U}$ to discover procedures on her own or to form a robust mental model of the system.
- Optimize messages for speed of presentation and/or comprehension, if necessary at the expense of other criteria.
  For example, synthesized speech could be played at a faster rate, even though it might sound less pleasant and require more effort to understand.

3 Ways of Recognizing Resource Limitations

Given that there appears to be some potential benefit to the automatic recognition of a user's resource limitations, on the basis of what evidence can a system achieve such recognition?

3.1 Recognizing Likely Causes of Resource Limitations

A system may be able to recognize factors that tend to give rise to resource limitations in users. Any evidence that suggests the presence of such a factor constitutes indirect evidence for the corresponding resource limitation. Table 1 gives some examples of the many possibilities.

3.2 Physiological Indicators

Within engineering psychology, there is a long tradition of research on physiological measures of cognitive load (see, e.g., [11, 12]). Such measures have mostly been applied in laboratory or field studies, but there is some potential for using them for on-line recognition of and adaptation to cognitive load. Two relatively promising measures can serve as examples.

Table 1 Examples of ways in which an adaptive system might obtain information about causes of a user's resource limitations

Cause of the resource limitation	Evidence of the cause that may be accessible to the adaptive system
Cognitive load	
Difficult driving situation	Information from navigation system
Use of a cognitively demanding interactive application	Information about applications currently being used by *U*
Distracting noise and/or events in the environment	Sensing of the environment through microphones or cameras
Time pressure	
Requirement for fast task completion imposed by the environment (e.g., flight for which boarding is about to close)	*S*'s access to information about environment-imposed constraints (e.g., boarding schedules)
Requirement for fast response imposed by *S* itself (e.g., instruction by *S* to perform a given action quickly)	*S*'s access to its own processing history

3.2.1 Heart Rate Variability

Heart rate variability (see, e.g., [13]) tends to decrease with increasing overall mental workload. In a study somewhat similar in spirit to the one to be described in Sect. 6, Rowe et al. [13] investigated the potential of heart rate variability to serve as in index of cognitive load, not only for the purpose of studying the workload induced by a given system but also for the purpose of allowing automatic adaptation. While this study did not yet yield clear conclusions about the value of heart rate variability for supporting on-line adaptation, they did suggest that further investigation of this possibility is warranted. Because of the need to attach electrodes to the user's body, heart rate variability does not fit especially naturally into the scenarios of mobile conversational interfaces; but perhaps ultimately the necessary sensors can be worn in an unobtrusive way and transmit data to a mobile device.

3.2.2 Pupil Diameter

The diameter of a person's pupil has likewise been shown to vary systematically as a function of mental load – although it is also strongly affected by other factors, such as ambient illumination and the distance of objects being fixated (see, e.g., [14]). These other factors would be especially problematic with mobile systems. Pupil diameter can be measured with eye-tracking equipment. With stationary system use, a remote eye tracker can be used that does not have to be attached to the user's head – although the user is required to sit relatively still. For mobile use, a head-mounted eye tracker is required; for the time being, therefore, this type of measurement must be restricted to research studies, as opposed to normal system use. As is the case with heart rate variability, studies are required to determine whether and in what situations this type of information can play a useful role in a system that adapts to a user's resource limitations.

A study conducted within READY illustrated that success is not guaranteed even in apparently optimal circumstances: In an experiment, Schultheis found no difference in the pupil diameter of subjects when they were reading very easy vs. very difficult texts on a computer screen (see, e.g., [15, 16]). A similar negative result was obtained by Iqbal et al. [17] on a similar reading task, but these same authors obtained good accuracy results on different types of tasks.

3.2.3 Other Indices

Other measures, which seem to have less immediate promise for use in mobile systems, include those that concern aspects of brain activity (for which, for example, Schultheis found some promising results in the experiment just mentioned; see also [18] for more recent and more promising results) and respiratory activity.

3.2.4 Comments

One general advantage of physiological measures is that in general a continuous stream of data is received without the need for the user to produce any particular behavior solely for diagnostic purposes. Some measures, such as heart rate variability and pupil diameter, respond quickly enough to changes in cognitive load to make on-line adaptation in principle feasible. A general drawback is the need for specialized sensors, which users may find uncomfortable or restrictive.

3.3 Evidence in the User's Behavior with the System

A different general class of evidence comprises information about the user's behavior in interacting with the system – for example, $\mathcal{U}$'s use of manual input devices or $\mathcal{U}$'s speech. One positive aspect of these types of evidence is that special sensing devices may be unnecessary, because the information enters $\mathcal{S}$ through the normal input channels. Moreover, $\mathcal{U}$'s input behavior (e.g., the fact that $\mathcal{U}$ is making manual input errors or producing disfluent speech) may be of importance in its own right – that is, a fact that $\mathcal{S}$ might adapt to or take into account in its processing.

3.3.1 Evidence in the User's Motor Behavior

Aspects of a user's motor behavior (e.g., tapping or dragging on a touchscreen with a stylus) could in principle reveal something about a user's resource limitations. A good deal of research has accumulated concerning features of motor behavior that typically arise under cognitive load and/or time pressure. Within the READY project, Lindmark [19] surveyed these relationships and suggested how they might be used for automatic recognition of resource limitations. For example, time pressure tends to lead to an increase in the stiffness of a person's limbs, which in turn tends to cause actions like tapping on the screen to be performed with relatively high force [20]; accordingly, when a given user employs more than the usual amount of force, this fact can be seen as suggestive evidence of time pressure. Cognitive load tends to

increase the likelihood of expert slips (e.g., forgetting to perform an intended action or tapping on an icon that looks similar to the intended one; cf. [10]); if the system can recognize such an error as having been made – in general not a trivial task – it can use the error as evidence that suggests cognitive load. Some behaviors (such as the two just mentioned as examples) are made more likely by either cognitive load or time pressure. Therefore, any mechanism for interpreting such evidence will have to have some appropriate mechanism for adjusting its hypotheses concerning both of these resource limitations on the basis of the same evidence. Although the emphasis in the present chapter is not on inference mechanisms, one possible such mechanism will be discussed in connection with the analyses in Sect. 7 and 8.

3.3.2 Evidence in the User's Speech

With conversational interfaces, an especially natural type of indicator of resource limitations comprises features of the user's speech. Because $\mathcal{S}$ needs to process $\mathcal{U}$'s speech anyway, there must already exist some type of microphone for sensing the speech and some software for analyzing it. Therefore, as with motor indicators, in the best case the only further requirements concern software for identifying and interpreting the indicators. The prospects for recognizing resource limitations on the basis of this type of indicator will be examined in detail starting in Sect. 4.

4 Experiments: Introduction

As was argued in 3.3.2, features of a user's speech appear in several respects to be a promising source of information about a user's cognitive resource limitations. But an obvious first question is: Is there enough information available in a user's speech to support a reasonably reliable recognition of these resource limitations?

4.1 Earlier Research on Speech Indicators

Before initiating a time-consuming experimental study, we surveyed previously conducted studies of relations between cognitive load or time pressure and features of speech.[2]

4.1.1 Distinction from Other Topics

The idea of making inferences about a speaker on the basis of features of their speech is by no means new. One topic of high practical importance is the recognition of emotion on the basis of speech (see, e.g., [32]). Part of this literature focuses on the effects of stress (see, e.g., [33]). Stress is related to cognitive load and time pressure, in that these resource limitations can be both causes and consequences

[2] Since this survey was made in 1998, it covered work through the late 1990s.

of stress. But there are also essential aspects of the concept of *stress* that are not necessarily associated with cognitive load or time pressure: physiological arousal and stressors such as noise or high acceleration (cf. [1], Chap. 12). We believe that it can be important to be able to adapt to cognitive load or time pressure even when these factors are not present – for example, when the user is performing two tasks at once and would like to proceed quickly but is not especially concerned about the consequences of failure. We therefore focus here on previous studies that did not involve especially stressful situations. (A much more detailed and comprehensive analysis of studies like these is given in [34].)

4.1.2 Effects of Cognitive Load

With regard to cognitive load, a number of features of speech have been investigated in multiple studies; hence it is possible to draw some fairly general conclusions concerning their dependence on cognitive load. Table 2 summarizes the most important of these indicators.

4.1.3 Effects of Time Pressure

Perhaps surprisingly, the number of results that can be extracted from previous studies concerning the effects of time pressure is much smaller than the number for cognitive load. One of the more obvious hypotheses is that people speak more quickly under time pressure. This hypothesis was confirmed in a study by Kelley

Table 2 Overview of the most important indicators of cognitive load found in some early studies

Indicator	Direction[a]	Tally[b]	Example study
Output rate			
Articulation rate	−	7/7	[21]
Speech rate	−	7/7	[22]
Pauses			
Onset latency (duration)	+/(−)	9/11	[23]
Silent pauses (number)	+	4/5	[24], Experiments 1 and 2
Silent pauses (duration, all)	+	6/8	[25]
Silent pauses (duration, intraphrasal only)	+	2/2	[26]
Filled pauses (number)	+	4/6	[27]
Filled pauses (duration)	+	1/2	[28]
Indicators involving output quality			
Repetitions (number)	+	5/6	[29], Experiment 2
Sentence fragments (number)	+	4/5	[24], Experiment 2
False starts (number)	+	2/4	[30]
Self-corrections (number)[c]	+, −, 0	2, 1, 4	[31]

[a] "+" means that the measure was generally found to increase under conditions of high cognitive load; "−" means the opposite.
[b] "m/n" means that of n relevant studies, m found the tendency indicated in the second column. (In most cases the tendency was statistically significant.)
[c] Results concerning self-corrections show an inconsistent pattern.

and Stone [35], and a study by Marx [36] showed a marginal tendency of the same sort. This same study by Marx revealed a statistically significantly greater tendency of speakers who had been put under time pressure to repeat parts of utterances.

5 Experimental Method

5.1 Purpose of Experiments

The goals of our two experiments were (a) to fill the gap in knowledge concerning the impact of time pressure on features of speech; (b) to examine within a single setting a large number of features that had previously mostly been studied separately; and (c) to obtain raw data that could be used to determine how well cognitive load and time pressure can be recognized on the basis of speech.

We required some way of capturing users' speech while they are subject to known resource limitations. In principle it would be possible to capture the speech in fairly natural conditions, if we could confidently assess the resource limitations in these conditions. Healy and Picard [37] applied this strategy in their study of physiological assessment of driver stress: Subjects were required to drive along a route that included a number of events which had predictable stress levels.

We chose an experimental setting for our studies, so as to be able to exert greater control over both the independent variables and the nature of the speech utterances.

We conducted two experiments, separated by about 1 year in time; Experiment 2 can be seen as a replication and extension of Experiment 1. For concreteness, Experiment 1 will be described separately first.

5.2 Method for Experiment 1

5.2.1 Materials

The experimental environment simulated a situation in which a user is walking through a crowded airport terminal while asking questions to a mobile assistance system via speech (see Fig. 2). In each of 80 trials, a picture appeared in the upper right-hand corner of the screen. On the basis of each picture, the subject was to ask a question, after motivating it with an introductory sentence. For example, for the picture shown in Fig. 2, a subject might say "I'm getting thirsty. Is there … will it be possible to get a beer on the plane?".

5.2.2 Design

Two independent variables were manipulated orthogonally:

- NAVIGATION: whether or not the subject was required to move an icon on the screen through the depicted terminal to an assigned destination by pressing arrow keys, while avoiding obstacles and remembering a gate number that comprised

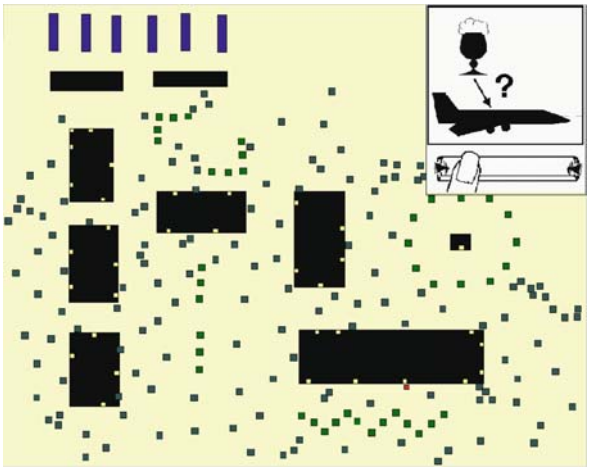


Fig. 2 Environment used in the experiments, with a typical pictorial stimulus

five digits and one letter. When navigation was not required, the subject could ignore the depicted terminal and concentrate on the generation of appropriate utterances in response to the pictures.
The navigation task was designed to induce the sort of cognitive load that would be induced by a nonverbal task performed by the user of a mobile system while interacting with the system. Walking around an airport would be one example of such a task; but there are of course differences between (a) walking in a real three-dimensional space and (b) moving an abstract figure within a two-dimensional computer screen. We do not refer to this condition as the "cognitive load" condition because it is not known to what extent the task actually induces cognitive load in any given subject.

- SPEECH TIME PRESSURE: whether the subject was induced by instructions and rewards (a) to finish each utterance as quickly as possible or (b) to create an especially clear and comprehensible utterance, without regard to time.
  More specifically, in the condition with time pressure, the subject was told that his speech would be interpreted by an experienced airport assistant who was in great demand because of her extensive knowledge. Utterances directed to this assistant were to be completed quickly, so that she could go on to assist other airport visitors. In the condition without time pressure, subjects were to direct their utterances to a new, inexperienced airport assistant. In this condition, nothing was said about time limitations; the emphasis was to be on ensuring that this assistant understood the utterances.

The instructions concerning SPEECH TIME PRESSURE make it almost inevitable for some differences in the speech of the subjects to appear as a function of this variable. Still, there are empirical questions concerning (a) the particular forms that the utterances take in the two conditions (e.g., whether, under SPEECH TIME

PRESSURE, subjects will articulate more quickly, use fewer words, and/or think less before starting to speak); and (b) whether the differences will be large enough to allow accurate discrimination between the two conditions.

We call this second variable SPEECH TIME PRESSURE to highlight its differences from other possible forms of time pressure. For example, if a person's goal is the quick completion of some larger task (e.g., getting to the departure gate), they may or may not try to save time by completing individual utterances quickly. But time pressure with regard to utterance completion can arise for various other reasons as well – for example, because of real or imagined time limitations on the part of the listener or system; because of a task that the user is performing that leaves only brief intervals free for speaking; or because of a high cost of utterances to the speaker, as in the case of an expensive communication channel. Any attempt to have a system adapt to SPEECH TIME PRESSURE in a given setting should take into account the likely reasons for this form of time pressure that might apply in that setting.

5.2.3 Procedure

After an extensive introduction to the scenario, the environment, and the 4 (2×2) conditions, each subject dealt with 4 blocks of trials, each block involving 20 pictures distributed over 4 destinations. Each block was presented in one of the 4 conditions, the order being varied across subjects according to standard procedures.

5.2.4 Subjects

The 32 subjects, students at Saarland University, were paid for their participation. An extra reward was given to one of the participants who most successfully followed the instructions regarding the time pressure manipulation.

5.2.5 Coding and Rating of Speech

Each of the 2,560 (32×80) utterances was transliterated and coded with respect to a wide range of features, including almost all of those that had been included in previous published studies. On the basis of the transliterations (minus the coding symbols), four independent raters sorted the stimulus pictures into five categories in terms of the complexity of the responses that they tended to call for. An aggregation of these ratings was later used to control for the different degrees of difficulty of the speech tasks invoked by the pictures.

In this chapter, we report results only for a subset of seven indicators which, on the basis of the results, seem most promising as indicators of cognitive load and/or time pressure[3]:

- NUMBER OF SYLLABLES: The number of syllables in the utterance.
- ARTICULATION RATE: The number of syllables articulated per second of speaking time, after elimination of the time for measurable silent pauses.

[3] Much more detailed reports covering all of the variables are given by Müller ([38], for Experiment 1) and by Kiefer ([39], for Experiment 2).

- SILENT PAUSES: The total duration of the silent pauses in the utterance, expressed relative to the length of the utterance in words (to take into account the fact that longer utterances offer more opportunities for pauses). In accordance with usual practice, a silent pause is defined as a silence within the utterance that lasts for at least 200 ms.
- FILLED PAUSES: The corresponding measure for filled pauses (e.g., "Uhh").
- HESITATIONS: The number of silences with a duration of less than 200 ms, again relative to the length of the utterance in words.
- ONSET LATENCY: The length of the time interval between the presentation of the pictorial stimulus and the first syllable spoken by the subject.
- DISFLUENCIES: The logical disjunction of several binary variables, each of which indexes one type of speech disfluency: self-corrections involving either syntax or content, false starts, or interrupting speech in the middle of a sentence or a word. Although each of these variables has been treated as a separate dependent variable in some previous studies, they are grouped together here because each phenomenon in question occurs too infrequently in our data to give rise to statistically reliable effects. (Filled and silent pauses, which may also be regarded as disfluencies, are not counted here, because they are treated as separate variables.)

5.3 Method for Experiment 2

The method for Experiment 2 was identical to that for Experiment 1, with one exception: During all of the time in which a subject was performing the experimental tasks, they heard through a headphone prerecorded loudspeaker announcements of the sort that travelers typically hear at airport terminals (concerning matters such as flight departures, gate changes, missing persons, and security warnings). These German-language announcements, which had been recorded at Frankfurt Airport, were arranged digitally so that there were only minimal pauses between announcements. For our present purposes, the function of these announcements was to add an additional source of cognitive load – one which, in contrast to the navigation task, seemed likely to interfere more directly with the process of speech production, because of its verbal nature.

Figure 3 gives a graphical overview of the eight specific conditions that were realized in the two experiments. Our focus will be on the effects that occurred within each experiment. Although it is of some theoretical interest to see how the announcements affected speech production, in the present chapter we will not pay much attention to a comparison of the results with and without announcements. One reason is that there is little practical interest attached to the question of whether a system can recognize, on the basis of a user's speech, whether that user is being distracted by irrelevant speech from the environment: If $\mathcal{U}$'s speech can be picked up by a microphone, then presumably the presence of ambient speech could be directly detected via the microphone as well. Also, from a methodological point of view, we must be cautious in interpreting specific differences between the results of Experiments 1 and 2: Even though considerable effort was made to replicate the

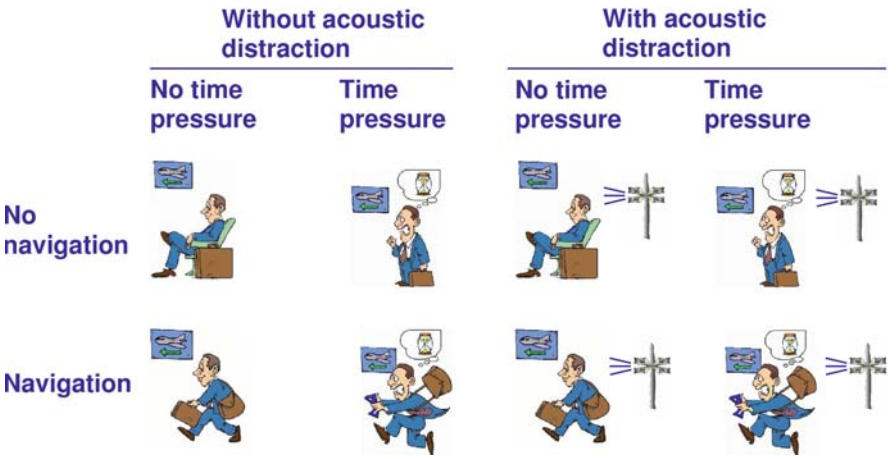


Fig. 3 Visualization of the eight conditions realized in Experiments 1 (*left*) and 2 (*right*). (In the experiments, the time pressure concerned specifically the time available to generate spoken input to the mobile system.)

method of Experiment 1, for practical reasons Experiment 2 was conducted by a different experimenter and the utterances were transliterated by a different researcher. Moreover, the subjects were not necessarily sampled from the same population. It is therefore most realistic to focus on the robust results which are found in both of the experiments despite the differences between them.

6 Experimental Results

6.1 Statistical Analyses

For each of the indicators analyzed here, a three-way analysis of variance (ANOVA) was conducted, with two within-subject variables (NAVIGATION and SPEECH TIME PRESSURE) and one between-subject variable (ANNOUNCEMENTS).[4] In accordance with the considerations just mentioned, we will interpret only the main effects of the within-subject variables and the interactions between them.

6.2 Number of Syllables

Figure 4 shows the means for the variable NUMBER OF SYLLABLES for each of the eight conditions. The ANOVA confirms that there is a highly significant main effect

[4] Before the ANOVAs were conducted, multivariate analyses of variance had been conducted with a view to ensuring against capitalizing on chance with the relatively large number of ANOVAs; these MANOVAs demonstrated that the interpretation of the ANOVAs reported here is justified.

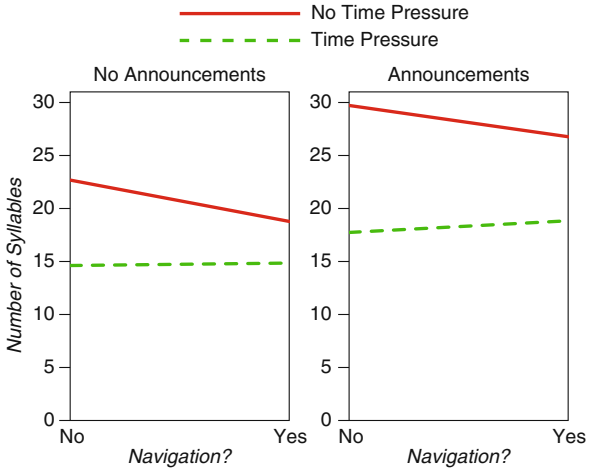


Fig. 4 Means for the variable NUMBER OF SYLLABLES in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

of SPEECH TIME PRESSURE ($F(1, 63) = 97.573$, $p < 0.001$): Not surprisingly, the instruction to finish each utterance quickly led to a much smaller number of syllables per utterance.

Somewhat less obviously, the requirement to navigate led to somewhat shorter utterances ($F(1, 63) = 8.295$, $p < 0.01$). Although there is no significant interaction between the two independent variables, the graphs suggest, plausibly, that the difference arises mainly in the condition without time pressure, in which the subjects were less ambitious with regard to the goal of producing unambiguous, high-quality utterances. When they were under time pressure, they were trying to keep their utterances short even when not navigating, so there was little room for the navigation task to cause further reduction in their length.

The results concerning NUMBER OF SYLLABLES are novel for the simple reason that previous studies have not in general included utterance length as a dependent variable. A likely reason for this omission is that utterance length has diagnostic significance only relative to a particular speech task: The fact that a user has produced a 15-syllable utterance in itself says little about her cognitive state; but if we know that the utterance was produced as an answer to a straightforward yes/no question, it may be significant. We will see in Sect. 7.1 how the properties of the current speech task can be taken into account in the interpretation of speech indicators.

6.3 Articulation Rate

As can be seen in Fig. 5, on the average subjects produced more syllables per second when they were under time pressure than when they were not ($F(1, 63) = 47.726$, $p < 0.001$). Though this result is intuitively plausible, it is not

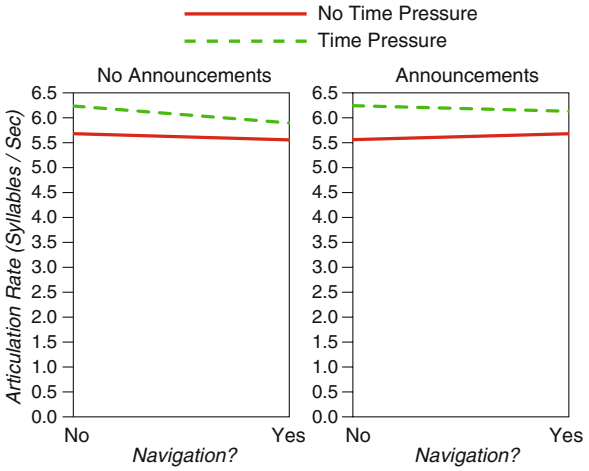


Fig. 5 Means for the variable ARTICULATION RATE in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

logically necessary, given that there are other ways of coping with time pressure (cf. Sect. 4.1.3). There is also a tendency to articulate less quickly when navigating (see the slope of the two lines; $F(1, 63) = 4.355$, $p < 0.05$), as has been reported in a number of previous studies (cf. Table 2). This effect is stronger under time pressure; this interaction ($F(1, 63) = 5.565$, $p < 0.05$) is understandable in that, under time pressure, subjects are articulating relatively fast, so there is more room for them to slow down.

The fact that the two main effects and the interaction are statistically significant, even though the differences involving ARTICULATION RATE do not appear visually striking in the graphs, testifies to the precision and sensitivity of ARTICULATION RATE as an index.

6.4 Silent Pauses

The results for SILENT PAUSES (Fig. 6) are complex. It is easily understandable that there is a highly significant main effect of SPEECH TIME PRESSURE ($F(1, 63) = 27.689$, $p < 0.001$): Without such pressure, subjects have no motivation to save time by avoiding pauses; perhaps even more importantly, they are motivated to produce high-quality utterances, which presumably tend to call for more careful planning, which can be accomplished during pauses. In particular, we have already seen (Fig. 4) that utterances produced without time pressure tend to be considerably longer; and as was shown by Oviatt [40], longer utterances tend to be associated with a relatively high number of disfluencies such as silent pauses.

Regarding the effects of NAVIGATION, previous studies (cf. Table 2) had shown that a concurrent task tends to increase the number and/or length of silent

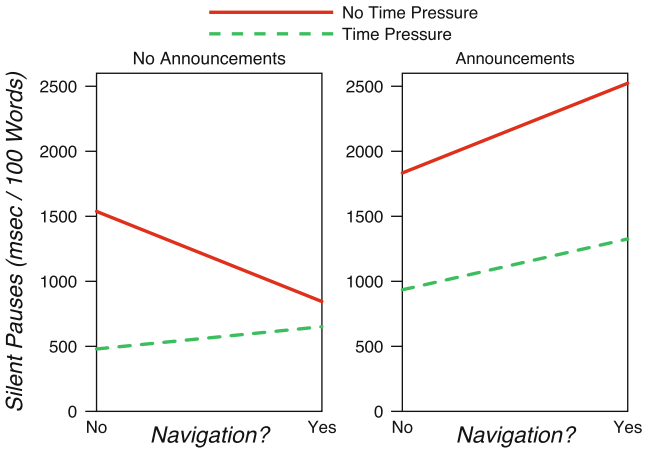


Fig. 6 Means for the variable SILENT PAUSES in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

pauses – plausibly enough, since a concurrent task demands the subjects' attention at least intermittently. This pattern is in fact seen in the upward slope of three of the four lines in Fig. 6. The reason why there is no significant overall main effect of NAVIGATION is that a sharp decrease occurs in Experiment 1 when there is no time pressure. This decline is understandable when we recall that, without time pressure, the need to navigate leads to shorter utterances (Fig. 4). In other words, subjects' adaptation to the navigation task proves more important in this case than the tendency of this task to increase cognitive load.

This specific result reminds us of a general point that is often emphasized in research on the effects of resource limitations on behavior (see, e.g., [1], Chap. 11; [41]). Resource limitations do not in general have a direct and unavoidable impact on performance; typically, a person has some freedom to decide how to deal with them.

6.5 Filled Pauses

With the indicator FILLED PAUSES (Fig. 7), the most striking difference between the two experiments appears. In Experiment 1 we see an effect that had been found in previous studies (cf. Table 2): an increase in filled pauses when a concurrent task is added. With the addition of the loudspeaker announcements in Experiment 2, this relatively subtle effect is reduced as the total duration of filled pauses increases by a factor of about 3; overall, there is no significant main effect of NAVIGATION. Although it is plausible that subjects generate more filled pauses in order to block out the distracting loudspeaker announcements, we should not attach much weight to this difference between the experiments, for the reasons given in Sect. 5.3.

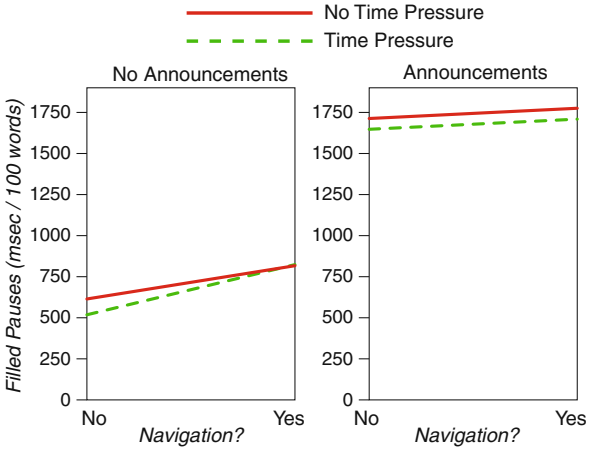


Fig. 7 Means for the variable FILLED PAUSES in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

6.6 Hesitations

The very short pauses counted by the variable HESITATIONS (Fig. 8) occur significantly less frequently when the subject is navigating ($F(1, 63) = 8.407$, $p < 0.01$); a possible explanation for this phenomenon is in terms of the reduction in the complexity of utterances when the subject is navigating (cf. Sect. 6.4). This result is novel in that virtually no previous studies have looked at hesitations as a dependent variable. The apparent effect of time pressure in the graphs is not statistically reliable, but note that it would be consistent with the results for SILENT PAUSES (Sect. 6.4).

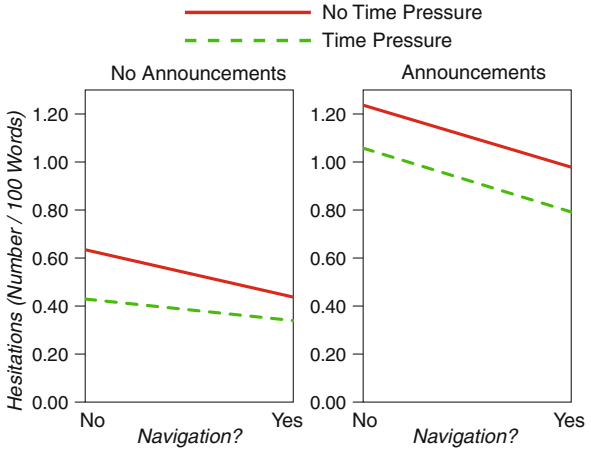


Fig. 8 Means for the variable HESITATIONS in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

6.7 Onset Latency

Regarding ONSET LATENCY (Fig. 9), we see a highly significant tendency for subjects to begin with the production of their utterance sooner when they have been instructed to get finished with the utterance quickly ($F(1, 63) = 95.841$, $p < 0.001$). In addition to the obvious explanation that they are simply following instructions, this effect may be due in part to the lower complexity of the utterances produced under time pressure (cf. Sect. 6.2), which reduces the amount of planning required. The tendency (suggested by the lack of parallelism in the lines of each graph) for ONSET LATENCY to be affected more by NAVIGATION when there is SPEECH TIME PRESSURE is confirmed by a significant statistical interaction between the two independent variables ($F(1, 63) = 8.079$, $p < 0.05$). The positive impact of cognitive load on onset latency that was found in many previous studies (see Table 2) is not found here to a statistically significant degree, although there is a visible tendency in that direction.

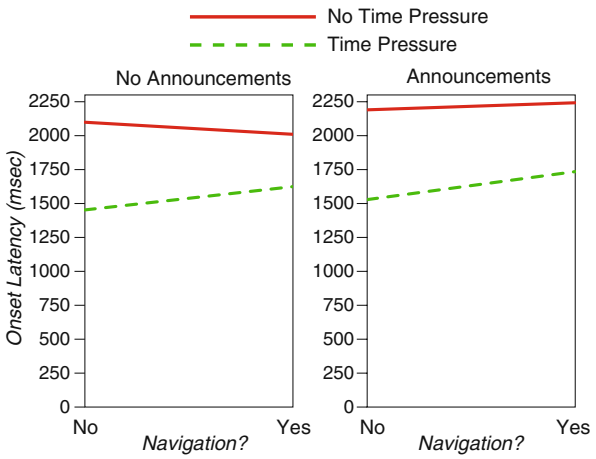


Fig. 9 Means for the variable ONSET LATENCY in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

6.8 Disfluencies

Although each of the specific types of disfluency summarized by the variable DISFLUENCIES occurs too infrequently to yield statistically significant differences as a function of the independent variables, a robust tendency does appear for the disjunction of the specific variables: As can be seen in Fig. 10, DISFLUENCIES increase when the subject is required to navigate ($F(1, 63) = 8.403$, $p < 0.01$, as was shown in previous studies (cf. Table 2). The other tendency that is apparent in the figure – for disfluencies to increase when there is no time pressure – is not statistically reliable in these data, though it would be consistent with the greater complexity of utterances generated when there is no time pressure (cf. [40]).

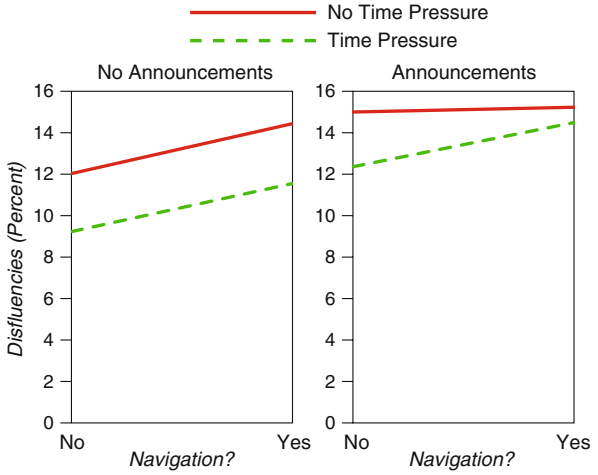


Fig. 10 Means for the variable DISFLUENCIES in the four conditions of Experiment 1 (*left*) and Experiment 2 (*right*)

6.9 Discussion

We have seen that, with the exception of FILLED PAUSES, each of the dependent variables discussed here shows one statistically reliable effect of time pressure and/or the navigation task. As was mentioned above, some of these results replicate and extend findings from previous experimental research, while others yield new information – especially those that concern the independent variable of SPEECH TIME PRESSURE and its interactions with the presence of a concurrent task.

Taken together, these results suggest that observation of these variables in a person's speech might allow a system to infer that person's current resource limitations. But the question of the extent to which such recognition is possible is not directly addressed by the conventional analyses that we have presented so far: A statistically significant result in an ANOVA shows that the result is unlikely to have occurred because of chance alone, but it does not guarantee that the dependent variable in question will have diagnostic value. To determine the prospects for recognizing resource limitations, we will apply quite different methods in the following two sections.

7 Learning of User Models

If we want to create a system that recognizes the resource limitations of its users on the basis of their speech, we need to take two main steps:

1. Use machine learning methods to create some sort of model relating resource limitations to speech indicators, using data such as those of these experiments (see the rest of this section).

2. Apply this model to the data of each user, using the features of their speech as evidence (Sect. 8).

7.1 Bayesian Network Structure

Regarding Step 1: There exists a great variety of machine learning techniques for classifying cases on the basis of their features, including support vector machines, neural networks, decision trees, and case-based reasoning.[5] A system that aims to recognize dynamically changing resource limitations imposes the following requirements on its learning and inference methods:

- The method should make it possible to interpret evidence from qualitatively different sources (cf. Sect. 3), ranging from likely causes of resource limitations to various types of indicator.
- The method should do justice to the fact that, while resource limitations change over time, the cognitive state of a user at any one moment will in most cases be similar to his or her state at the previous moment.
- The modeling method should yield a more or less interpretable model: Especially when several qualitatively different types of evidence are being used, it should be possible, by inspection of the model, to understand their relationships to one another (cf. [45]). Otherwise, it may be difficult to adapt the method to scenarios that involve different types of evidence.
- It should be possible to acquire a model of each individual user, so as to be able to take into account individual differences in the ways in which resource limitations are reflected in speech. But user model acquisition should also be able to take advantage of data acquired from users other than the current user, so that learning does not have to begin from scratch with each new user (cf. [46]).

Among the learning and inference techniques that best fit this combination of requirements are those that are associated with Bayesian networks (BNs).[6]

The BN structure employed in the present study is illustrated in Fig. 11. (The nodes in the lower box labeled TIME SLICE 2 can be ignored for the moment.) We will first consider its qualitative structure; the quantitative modeling of the relationships among the variables represented will be discussed below.

The three nodes NAVIGATION, SPEECH TIME PRESSURE, and ANNOUNCEMENTS on the left correspond to the three main independent variables of the experiments. The node DIFFICULTY OF SPEECH TASK refers to the rated complexity of the speech task created by the stimulus picture (cf. Sect. 5). Each of these nodes represents a variable that can be seen as influencing the values of the seven dependent

[5] For general treatments of machine learning techniques, see [42, 43]. Applications of such techniques to the modeling of computer users are discussed in [44].

[6] The technical aspects of the use of Bayesian networks in the READY project, with a focus on the learning of BNs, are discussed in the chapter by Wittig, comparison of Machine Learning Techniques for Bayesian Networks for User-Adaptive Systems in this volume.

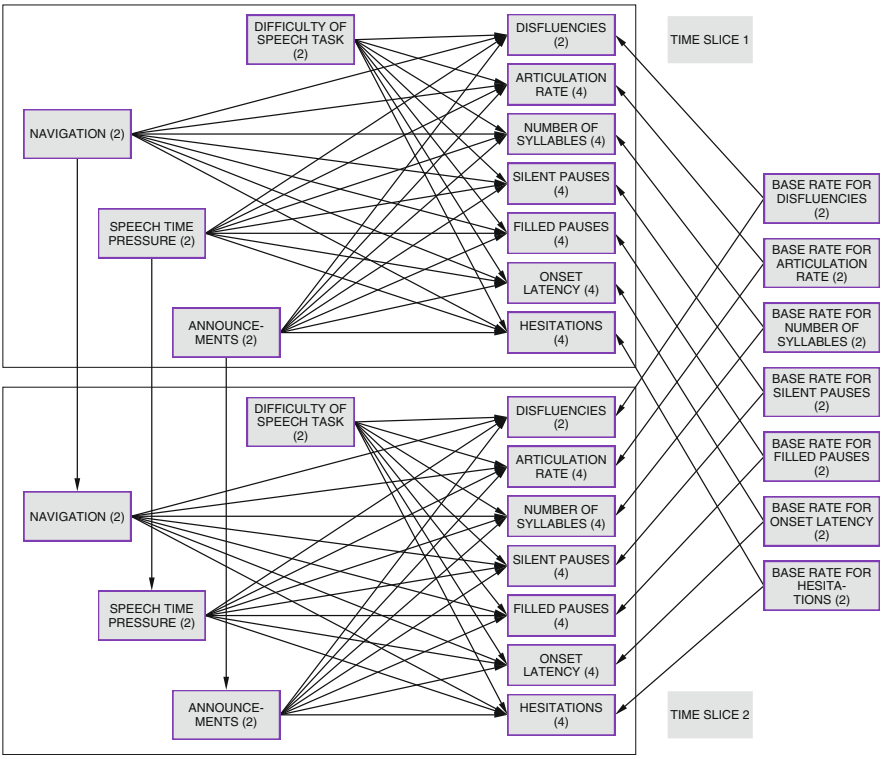


Fig. 11 Structure of the dynamic Bayesian network used in the evaluation of recognition accuracy. (Nodes within the two large boxes correspond to temporary variables that index features of the current utterance. Each number in parentheses shows the number of discrete states for the variable in question.)

(indicator) variables that were analyzed in Sect. 6; these variables are represented by the seven nodes on the right within the box for TIME SLICE 1. Further influences on the indicator variables are represented by the seven nodes on the far right in the figure, which correspond to individual base rates for the seven indicator variables. They are introduced to take into account individual differences in the overall level of the indicator variables. The value of each such variable is constant for each $\mathcal{U}$: It is simply computed as the mean value of the variable in question for the entire experiment.

The BN structure in the figure shows a rather drastic simplification of the causal relationships that actually exist between the variables in question. For example, the absence of links among the base rate variables implies that these variables are statistically independent. In addition to being implausible, this assumption was shown to be false by our own factor analyses and applications of algorithms for learning BN structures from data. Nonetheless, this simplified model was found to perform

better at the task of recognizing a speaker's time pressure and cognitive load than did more complex models that took into account the statistical dependencies.[7]

Our question in the evaluation study will be: If a user $\mathcal{U}$ produces a sequence of utterances in a given experimental condition, how well can a system $\mathcal{S}$ recognize what condition the user was in? Therefore, the variables NAVIGATION and SPEECH TIME PRESSURE can be viewed here as *static* variables whose value does not change over time. The seven base rate variables are also static. By contrast, each of the variables inside the boxes labeled TIME SLICE 1 and TIME SLICE 2 refers to an aspect of just one utterance. Hence corresponding *temporary* nodes need to be created for each utterance. We are therefore dealing with a *dynamic Bayesian network* (DBN) that comprises a series of *time slices*.[8]

7.2 Quantitative Parameters

In a BN such as the one used here, which does not include continuous variables, each variable has two or more discrete *states*, or possible values. For example, for NAVIGATION, the two states are "Navigation" and "No navigation". For the base rate variable BASE RATE FOR NUMBER OF SYLLABLES, each state corresponds to one of four ranges of numbers of syllables.

For each *root node* (i.e., a node that has no links directed at it), the system's initial expectation about the value of the variable in question is represented by a vector of probabilities that represents a probability distribution. For example, for each of the nodes SPEECH TIME PRESSURE, NAVIGATION, and ANNOUNCEMENTS, the probabilities are simply <0.50, 0.50>, reflecting the fact that each value of each of these variables occurred equally often in the experiments. For each of the base rate nodes, the probability vector reflects the empirically determined distribution of the base rate in question in the group of subjects in these experiments.

For each node that is not a root node, a *conditional probability table* (CPT) represents the system's assumptions about how the value of the variable is related to the values of its *parent variables* (corresponding to the nodes with links that point to it). For example, each probability in the CPT for DISFLUENCIES represents the likelihood that a disfluency will occur (or not occur) in an utterance, given particular values of the parent variables SPEECH TIME PRESSURE, NAVIGATION, ANNOUNCEMENTS, DIFFICULTY OF SPEECH TASK, and BASE RATE FOR DISFLUENCIES.

A BN makes probabilistic inferences when it is evaluated: Typically, one or more variables in the BN are *instantiated*; that is, the probability distribution representing the system's belief about the value of such a variable is replaced by a probability

[7] A possible reason is that in the more complex models the estimates of some probabilities in the learned BN are less accurate because they are based on relatively few observations.

[8] An explanation of the general principles of dynamic Bayesian networks can be found, for example, in Chap. 17 of [47]. A discussion with regard to user modeling of the sort done here is given in [48].

distribution which expresses certainty that one particular value is realized. Then the BN is reevaluated; typically the system's beliefs about some of the uninstantiated variables are updated to be consistent with the new information provided by the instantiations.

7.3 Learning the Quantitative Parameters

Although we specified the structure of the BN shown in Fig. 11 by hand, the probabilities need to be learned empirically. Such learning is quite straightforward in a BN (such as this one) that includes only observable variables: In accordance with the usual maximum-likelihood method (see, e.g., [49]), the estimate of each (conditional) probability is computed simply in terms of the (relative) frequencies in the data.[9]

Since we want to test a learned BN model with the data of a given user $\mathcal{U}$, we must not include $\mathcal{U}$'s data in the data that are used for the learning of the corresponding BN. Accordingly, we learned for each $\mathcal{U}$ the conditional probability tables for a separate BN using the data from the other 63 subjects. The learned BN has the structure shown in Fig. 11 minus the nodes shown for TIME SLICE 2; the CPTs for the temporary variables within each time slice are the same as the ones learned for TIME SLICE 1.

8 Evaluation of the User Models

8.1 Procedure

The basic idea of the evaluation of the learned models can be explained with reference to Fig. 3: Given the behavior of a subject in one of the eight experimental conditions, our system will try to infer which condition the subject was in when he or she produced that behavior. More specifically, when asking the system to assess the probability that $\mathcal{U}$ was under time pressure, we will tell the system whether $\mathcal{U}$ was navigating and whether $\mathcal{U}$ was distracted by loudspeaker announcements. Similarly, when asking the system to assess the probability that $\mathcal{U}$ was navigating, we will specify the true values of the other two independent variables. (We will not report on tests of how well $\mathcal{S}$ can discriminate between the presence and the absence of ANNOUNCEMENTS, for the reasons given in Sect. 5.3, except to note in passing that the results are roughly comparable to those reported below for the recognition of NAVIGATION.)

More formally, the procedure for evaluating a learned BN is given in Table 3.

---

[9] The learning of BNs in much more complex settings is discussed in the chapter by Wittig, Comparison of Machine Learning Techniques for Bayesian Networks for User-Adaptive Systems in this volume.

Table 3 Procedure used in evaluating the accuracy with which a learned Bayesian network assesses the value of the variable SPEECH TIME PRESSURE for a given user. (The procedure is identical when the value of NAVIGATION is to be assessed, except that the roles of T and N are interchanged.)

Relevant variables and their values

- A user U
- Values t, n, and a of the Boolean variables T (Speech Time Pressure), N (Navigation), and A (Announcements)

Task
Infer the value of T on the basis of indicators in U's speech

Preparation of the test data
Select the 20 observations for U in which $T = t$, $N = n$, and $a = A$, in the order in which they occurred in the experiment in question

Evaluating recognition accuracy
Initialize the model:

1. Create the first time slice of the BN for U
2. Instantiate each of the individual base rate variables with its true value for U
3. Also instantiate N and A with their true values n and a, but leave the variable T (whose value is to be inferred) uninstantiated

For each observation O in the set of observations for U:

1. In the newest time slice of the BN, derive a belief about T:
   - Instantiate all of the temporary variables for this time slice with their values in O
   - Evaluate the BN to arrive at a belief regarding T
   - Note the probability assigned at this point to the true value t of T
2. Add a new time slice to the dynamic BN to prepare for the next observation

8.2 Results

Because of the differences between Experiments 1 and 2 (cf. Sect. 5.3), in Fig. 12 the results of the modeling evaluation are shown separately for each of the two experiments. Each curve is the result of averaging 32 curves, one for each subject in the experiment in question.[10]

8.2.1 Recognizing Time Pressure

Looking first at the results for recognizing SPEECH TIME PRESSURE (left-hand graphs), we see that the BNs are on the whole rather successful: The average probability assigned to the actual current condition rises sharply during the first few observations. Note that in each experiment, recognition of SPEECH TIME PRES-

[10] The results for individual subjects are much less smooth than these aggregated results: The individual curves often show sharp jumps and extreme values.

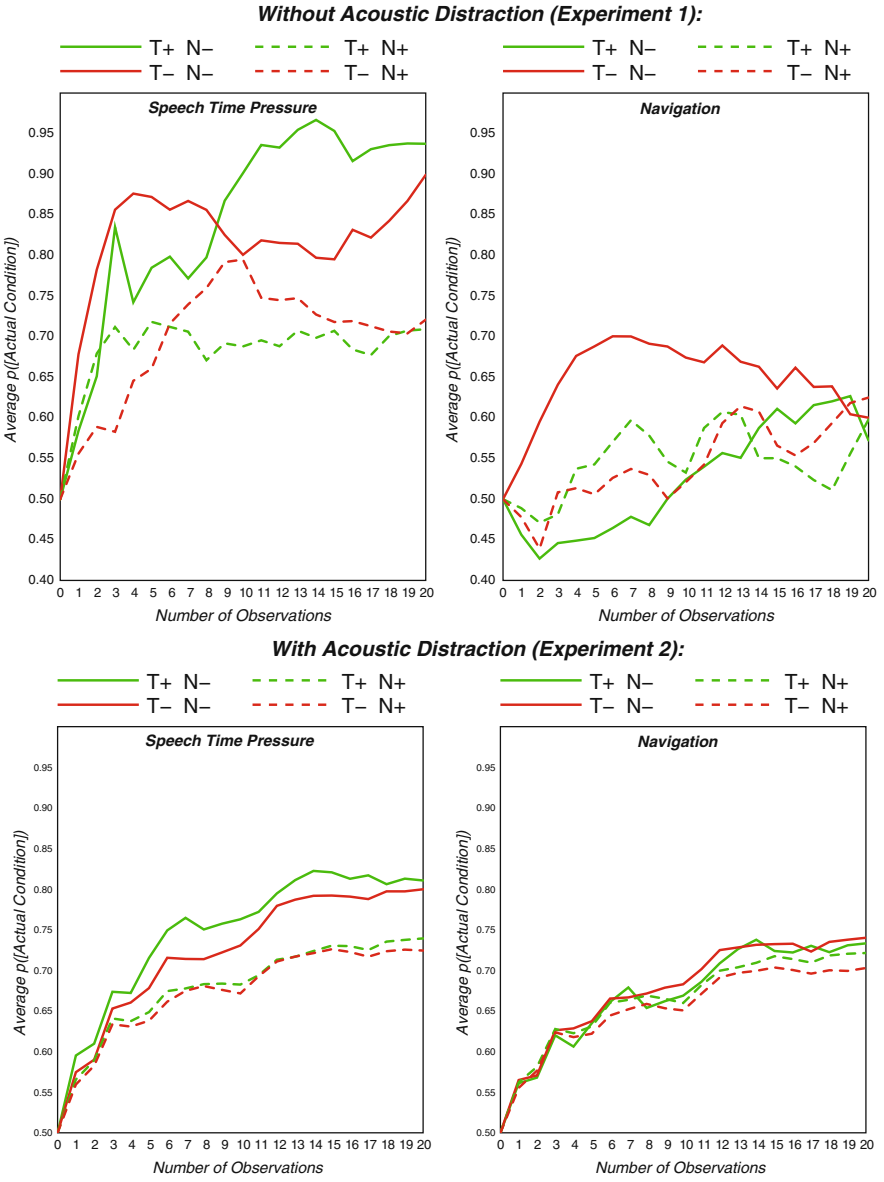


Fig. 12 Accuracy of the learned dynamic Bayesian networks in inferring the correct value of SPEECH TIME PRESSURE ("T", *left*) and NAVIGATION ("N", *right*) in Experiment 1 (*above*) and Experiment 2 (*below*). (Each curve shows the aggregated results for one combination of values of the variables SPEECH TIME PRESSURE, NAVIGATION, and ANNOUNCEMENTS. In each curve, the point for the *i*th observation shows the average probability which the Bayesian network assigned to the subject's actual condition after processing the first *i* observations.)

SURE is easier when there is no navigation task.[11] This result is understandable in the light of the conventional analyses discussed in Sect. 6: On the whole, the effects of time pressure were somewhat greater when there was no navigation task (i.e., the lines tended to be farther apart on the left-hand sides of the graphs), since in that condition speakers were able to respond more sensitively to the time pressure (or lack of it).

8.2.2 Recognizing Navigation

In Experiment 2, the results for recognition of NAVIGATION are consistent over the four conditions: After several observations, the system on the average assigns a probability of roughly 0.65 to the correct condition. The fact that this probability never rises much above 0.70, even after 20 observations, shows that there is an inherent difficulty in discriminating between the presence and absence of NAVIGATION which cannot be overcome through the provision of a large number of observations.

In Experiment 1, the results are generally poorer, and they show rather strange variations between conditions and over time.[12] One way of understanding the better results for Experiment 2 is simply to note that the indicators shown in Fig. 6, 7, 8, 9 and 10 tend to occur to a greater extent in Experiment 2 (i.e., the lines in the right-hand graphs in these figures tend to be higher than those in the left-hand graphs). Since these indicators are on the whole low-frequency events, any increase in their frequency is likely to make recognition more accurate. It may be speculated that this overall difference in the frequency of indicators is due to the presence of loudspeaker announcements in Experiment 2, which push subjects closer to the limits of their processing capacity.

8.2.3 Dispensing with Individual Indicators

Especially when we consider the practical problem of measuring indicators automatically (see Sect. 8.3), it becomes interesting to consider which of the seven indicators might be dispensable on the grounds that they do not add significantly to the accuracy of recognition. We repeated the simulations summarized in Fig. 12 seven times, each time leaving out one of the seven indicators. Since it would be tedious and imprecise to examine seven further sets of four graphs similar to those shown in Fig. 12, we computed for each graph a single number that

[11] Since this statement applies to each of the observations 1–20 in each experiment, the difference is statistically reliable for each experiment with $p < 0.001$ by a sign test.

[12] As was mentioned in an earlier report on Experiment 1 [50], the results for the recognition of navigation are actually better if the system is *not* told whether $\mathcal{U}$ was under time pressure – perhaps because the BN then bases its assessment on a larger number of conditional probabilities and hence, indirectly, on a larger amount of data from other subjects. Overall, however, there is no systematic tendency for recognition to be better or worse when the system is told the value of the independent variable(s) that it is not trying to assess.

summarizes the success of recognition: the mean of the 80 probabilities shown in the four curves of the graph. The question then becomes: To what extent do these mean probabilities decline when one of the indicator variables is left out of consideration?

Table 4 shows the results. The indicator whose removal has the greatest impact is clearly NUMBER OF SYLLABLES. Each of the other indicators seems surprisingly dispensable; and in a few cases leaving an indicator out even improves recognition accuracy. As the final column shows, the sum of the changes that result from leaving individual indicators out is much smaller than the extent to which recognition exceeds the chance level of 50%. This fact shows that the contributions of the indicators are not simply additive: It may be possible to leave out one indicator without much loss of accuracy because the information that it contributes is largely supplied by other indicators; but it would not be advisable to leave out all or most of them.

The indicator that it would presumably be most practically useful to omit is DISFLUENCIES: Automatically recognizing linguistic phenomena such as self-corrections, false starts, and interrupted sentences is considerably more difficult than measuring (silent or filled) pauses and counting syllables, which is all that is required for the other indicators.[13] As Table 4 shows, the variable DISFLUENCIES adds at best negligible value, provided that the other indicators are available.

8.3 Discussion

One question concerns the extent to which the results concerning the recognition of resource limitations can be generalized to different (and more realistic) settings. Certainly the specific probabilities of correct recognition are dependent on features of the particular situation – witness the differences that arose even between these two very similar experiments. For our analyses, it was certainly helpful that the experimental situation was highly constrained. Moreover, it was important for the system to know the difficulty of the specific speech task that the user was performing. In an interactive system, the corresponding information would consist in expectations about the complexity of the utterance that the user is likely to produce in any given situation (for example, after a question about the user's desired destination).

In sum, much work remains to be done before features in a user's speech can be used for the recognition of the resource limitations of a real user of an interactive system; and even in the long run this possibility will probably be subject to various restrictions – for example, concerning the predictability of the speech produced by users.

---

[13] Portable hardware (with associated software) for detecting and analyzing pauses in speech is commercially available.

Table 4 Impact on recognition accuracy of leaving out of consideration each of the seven indicator variables. (Each number in the column "All indicators" is the mean of the 80 probabilities shown in the corresponding graph in Fig. 12, expressed as a percentage. Each number in a column to the right (except the rightmost column) shows the corresponding mean change in accuracy (as a percentage, but in absolute terms) when the simulation is performed without use of the indicator variable in question. The rightmost column shows the sum of these changes

	All indicators	Syllables	Articulation rate	Filled pauses	Hesitations	Onset latency	Disfluencies	Silent pauses	Sum of changes
Speech time pressure:									
Experiment 1	75.76	−6.13	0.00	−0.09	+0.07	−3.61	+0.36	+1.12	−8.29
Experiment 2	70.32	−5.86	−1.17	−1.21	−0.60	−0.18	−0.04	−0.04	−9.10
Navigation:									
Experiment 1	56.58	−1.86	−0.66	−2.02	−0.29	+1.39	+0.12	−0.99	−4.31
Experiment 2	66.51	−4.94	−1.11	−0.84	−0.54	−0.35	−0.03	+0.44	−7.39

9 Summary of Contributions and Remaining Work

One goal of the present chapter was to provide a framework for thinking about the prospects for adapting to a user's cognitive resource limitations in interactive systems in general and in mobile conversational systems in particular. We discussed why such adaptation might be worthwhile, what forms it might take, and how the resource limitations might be automatically assessed.

The more specific goal was to explore the prospects of exploiting the user's speech as a source of evidence for the recognition of resource limitations. One respect in which the two experiments presented differ from comparable previous experiments concerns the number of independent variables examined simultaneously: Whereas almost all previous studies had examined the effects of just one variable (usually cognitive load), our experiments orthogonally manipulated cognitive load and speech time pressure, as well as repeating the experiment with and without distraction from irrelevant speech. The nature of the manipulations makes the experiments somewhat more relevant to scenarios of mobile conversational interaction than previous experiments were. But the most important new contribution concerns the results on the diagnostic value of seven specific features of speech: The evaluation experiments show that these indicators together do permit a degree of recognition of time pressure and cognitive load that could be useful in some situations, and they indicate the effects of leaving out individual features that would be relatively hard to recognize automatically.

Any attempt to apply the ideas and results from this chapter in a particular application scenario will necessarily involve considerable further work and creativity. But we believe that the results presented here will be helpful as a starting point.

Acknowledgments The research described here was supported by the German Science Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Projects B2 (READY) and A2 (VEVIAG). Preparation of this manuscript was supported by the Province of Trento in its targeted research unit Prevolution (code PsychMM). The research benefited greatly from preparatory studies by André Berthold [34] and from advice by Werner Tack. Some results concerning Experiment 1 were described in a conference paper by Müller et al. [50].

References

1. Wickens, C.D., Hollands, J.G. Engineering Psychology and Human Performance (3rd edn). Upper Saddle River, NJ: Prentice Hall (2000).
2. Wierwille, W.W. Visual and manual demands of in-car controls and displays. In Peacock, B., Karwowski, W. (Eds), Automotive Ergonomics (pp. 299–320). London: Taylor and Francis (1993).
3. Bernsen, N.O., Dybkjær, L. Exploring natural interaction in the car. In Proceedings of the CLASS Workshop on Natural Interactivity and Intelligent Interactive Information Representation. Verona, Italy (2001).
4. Salvucci, D.D. Predicting the effects of in-car interfaces on driver behavior using a cognitive architecture. In Jacko, J.A., Sears, A., Beaudouin-Lafon, M., Jacob, R.J. (Eds.), Human

Factors in Computing Systems: CHI 2001 Conference Proceedings (pp. 120–127). New York: ACM (2001).
5. Sawhney, N., Schmandt, C. Nomadic Radio: Speech and audio interaction for contextual messaging in nomadic environments. ACM Transactions on Computer-Human Interaction, 7:353–383 (2000).
6. Horvitz, E. Principles of mixed-initiative user interfaces. In Williams, M.G., Altom, M.W., Ehrlich, K., Newman, W. (Eds.), Human Factors in Computing Systems: CHI 1999 Conference Proceedings (pp. 159–166). New York: ACM (1999).
7. Horvitz, E., Jacobs, A., Hovel, D. Attention-sensitive alerting. In Laskey, K.B., Prade, H., (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference (pp. 305–313). San Francisco: Morgan Kaufmann (1999).
8. Litman, D.J., Pan, S. Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction, 12(2–3):111–137 (2002).
9. Jameson, A., Großmann-Hutter, B., March, L., Rummer, R., Bohnenberger, T., Wittig, F. When actions have consequences: Empirically based decision making for intelligent user interfaces. Knowledge-Based Systems, 14:75–92 (2001).
10. Norman, D.A. Design rules based on analyses of human error. Communications of the ACM, 26:254–258 (1983).
11. Kramer, A.F. Physiological metrics of mental workload: A review of recent progress. In Damos, D.L. (Ed.), Multiple-Task Performance (pp. 279–328). London: Taylor and Francis (1991).
12. Wilson, G.F., Eggemeier, F.T. Psychophysiological assessment of workload in multi-task environments. In Damos, D.L. (Ed.), Multiple-Task Performance (pp. 329–360). London: Taylor and Francis (1991).
13. Rowe, D.W., Silbert, J., Irwin, D. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Karat, C.M., Lund, A., Coutaz, J., Karat, J. (Eds.), Human Factors in Computing Systems: CHI 1998 Conference Proceedings (pp. 480–487). New York: ACM (1998).
14. Granholm, E., Asarnow, R.F., Sarkin, A.J., Dykes, K.L. Pupillary responses index cognitive resource limitations (pp. 457–461). Psychophysiology, 33(4):(1996).
15. Schultheis, H. Pupillengröße und kognitive Belastung [Pupil size and cognitive load]. Master's thesis, Saarland University, Department of Psychology (2004).
16. Schultheis, H., Jameson, A. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In Nejdl, W., De Bra, P., (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems: Proceedings of AH 2004 (pp. 225–234). Berlin: Springer (2004).
17. Iqbal, S.T., Zheng, X.S., Bailey, B.P. Task-evoked pupillary response to mental workload in human-computer interaction. In Extended Abstracts for CHI 2004 (pp. 1477–1480). Vienna (2004).
18. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., Rao, R.P. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In Burnett, M., Costabile, M.F., Catarci, T., de Ruyter, B., Tan, D., Czerwinski, M., Lund, A. (Eds.), Human Factors in Computing Systems: CHI 2008 Conference Proceedings (pp. 835–844). New York: ACM (2008).
19. Lindmark, K. Interpreting symptoms of cognitive load and time pressure in manual input. Master's thesis, Department of Computer Science, Saarland University, Germany (2000).
20. van Galen, G.P., van Huygevoort, M. Error, stress and the role of neuromotor noise in space oriented behaviour. Biological Psychology, 51:151–171 (2000).
21. Lazarus-Mainka, G., Arnold, M. Implizite Strategien bei Doppeltätigkeit: Sprechen = Zuhören und Sortieren [Implicit strategies in dual tasks: Speaking = listening and sorting]. Zeitschrift für experimentelle und angewandte Psychologie, 34:286–300 (1987).
22. Kowal, S., O'Connell, D. Some temporal aspects of stories told while or after watching a film. Bulletin of the Psychonomic Society, 25:364–366 (1987).
23. Greene, J.O. Speech preparation processes and verbal fluency. Human Communication Research, 11:61–84 (1984).

24. Rummer, R. Kognitive Beanspruchung beim Sprechen [Cognitive load in speaking]. Weinheim, Germany: Beltz (1996).
25. Goldman-Eisler, F. Psycholinguistics: Experiments in Spontaneous Speech. London: Academic Press (1968).
26. Butterworth, B. Evidence from pauses in speech. In Butterworth, B. (Ed.) Language Production (pp. 155–176). New York, London: Academic Press (1980).
27. Wiese, R. Psycholinguistische Aspekte der Sprachproduktion [Psycholinguistic Aspects of Speech Production]. PhD thesis, Hamburg (1983).
28. Grosjean, F., Deschamps, A. Analyse des variables temporelles du français spontané [analysis of temporal variables in spontaneous French]. Phonetica, 28:191 (1973).
29. Deese, J. Pauses, prosody, and the demands of production in language. In Dechert, H.W., Raupach, M. (Eds.), Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler (pp. 69–84). The Hague: Mouton (1980).
30. Roßnagel, C. Übung und Hörerorientierung beim monologischen Instruieren: Zur Differenzierung einer Grundannahme [Practice and listener-orientation in the delivery of instruction monologs: Differentiation of a basic assumption]. Sprache & Kognition, 14:16–26 (1995).
31. Oviatt, S. Predicting spoken disfluencies during human–computer interaction. Computer Speech and Language, 9:19–35 (1995).
32. Banse, R., Scherer, K.R. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70(3):614–636 (1996).
33. Fernandez, R., Picard, R.W. Modeling drivers' speech under stress. In Proceedings of the ISCA Workshop on Speech and Emotions. Belfast (2000).
34. Berthold, A. Repräsentation und Verarbeitung sprachlicher Indikatoren für kognitive Ressourcenbeschränkungen [Representation and processing of linguistic indicators of cognitive resource limitations]. Master's thesis, Saarland University, Department of Computer Science (1998).
35. Kelly, K.R., Stone, G.L. Effects of time limits on the interview behaviour of Type A and B persons within a brief counseling interview. Journal of Counseling Psychology, 29:454–459 (1982).
36. Marx, E. Über die Wirkung von Zeitdruck auf Sprachproduktionsprozesse [The Effect of Time Pressure on Speech Production Processes]. PhD thesis, University of Münster, Germany (1984).
37. Healey, J., Picard, R. SmartCar: Detecting driver stress. In Proceedings of the Fifteenth International Conference on Pattern Recognition (pp. 4218–4221). Barcelona (2000).
38. Müller, C. Symptome von Zeitdruck und kognitiver Belastung in gesprochener Sprache: eine experimentelle Untersuchung [Symptoms of time pressure and cognitive load in speech: An experimental study]. Master's thesis, Department of Computational Linguistics, Saarland University, Germany (2001).
39. Kiefer, J. Auswirkungen von Ablenkung durch gehörte Sprache und eigene Handlungen auf die Sprachproduktion [Effects on speech production of distraction through overheard speech and one's own actions]. Master's thesis, Department of Psychology, Saarland University, Germany (2002).
40. Oviatt, S. Multimodal interactive maps: Designing for human performance. Human-Computer Interaction, 12:93–129 (1997).
41. Baber, C., Mellor, B. The effects of workload on speaking: Implications for the design of speech recognition systems. In Contemporary Ergonomics: Proceedings of the Annual Conference of the Ergonomics Society, 513–517 (1996).
42. Langley, P. Elements of Machine Learning. San Francisco: Morgan Kaufmann (1996).
43. Mitchell, T.M. Machine Learning. Boston: McGraw-Hill (1997).
44. Webb, G., Pazzani, M.J., Billsus, D. Machine learning for user modeling. User Modeling and User-Adapted Interaction, 11:19–29 (2001).
45. Wittig, F., Jameson, A. Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In Boutilier, C., Goldszmidt, M. (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (pp. 644–652). San Francisco: Morgan Kaufmann (2000).

46. Jameson, A., Wittig, F. Leveraging data about users in general in the learning of individual user models. In Nebel, B. (Ed.), Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (pp. 1185–1192). San Francisco: Morgan Kaufmann (2001).
47. Russell, S.J., Norvig, P. Artificial Intelligence: A Modern Approach (2nd edn). Englewood Cliffs, NJ: Prentice-Hall (2003).
48. Schäfer, R., Weyrath, T. Assessing temporally variable user properties with dynamic Bayesian networks. In Jameson, A., Paris, C., Tasso, C. (Eds.), User Modeling: Proceedings of the Sixth International Conference, UM97 (pp. 377–388). Vienna: Springer Wien New York (1997).
49. Buntine, W. A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 8:195–210 (1996).
50. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Bauer, M., Gmytrasiewicz, P., Vassileva, J. (Eds.), UM2001, User Modeling: Proceedings of the Eighth International Conference (pp. 24–33). Berlin: Springer (2001).

The Shopping Experience of Tomorrow: Human-Centered and Resource-Adaptive

Wolfgang Wahlster, Michael Feld, Patrick Gebhard, Dominikus Heckmann, Ralf Jung, Michael Kruppa, Michael Schmitz, Lübomira Spassova, and Rainer Wasinger

1 Introduction

What would the shopping experience of tomorrow look like? In this chapter we propose several human-centered and resource-adaptive ideas to this question. Throughout the whole chapter we explain our ideas with the recurrent theme of a shop that consists of instrumented shelves, public displays, audio systems, and mobile devices for each user. The shelves are fitted with RFID antennas and allow for sensing implicit user interactions with RFID-labeled objects, such as picking up a product or putting it back into the shelf. We will present the novel interaction paradigm of "Talking Objects", which involves multimodal interaction with instrumented objects, spanning the modalities of speech, gestures, sound and haptics. Imagine talking objects in shopping malls with which individuals or groups are able to interact. This means accomplishing shopping tasks by offering an intuitive interface to a complex environment. Furthermore, these talking objects will be associated with personalities by the means of controlling speech attributes and behavior. In addition to this anthropomorphism, we will provide these objects with the abilities to sense their state, e.g., whether they are in or outside the shelf, or whether a user is turning, squeezing, or shaking them. The novel concept of "Product Associated Displays" is a way of providing visual feedback to users interacting with physical objects in an instrumented shop. These projected public displays are created at locations that can be intuitively associated with the objects they show information about. Furthermore, a life-like character lives as a "Virtual Room Inhabitant" in our smart shop. The novel concept of "Personalized Ambient Audio Notification" describes a notification service that allows users to monitor information with less distraction of attendees in their surrounding. The ambient notification service works with personalized non-speech audio cues that can be embedded in aesthetic background music depending on the event and the current position of the user. Areas of applications are shops where employees can receive information (e.g., a cashier

W. Wahlster (✉)
DFKI GmbH and Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany
e-mail: wahlster@dfki.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,* Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_10,

is needed in the point of sale area) without arousing the customer's attention. At the same time the background soundscape has a comfortable effect on customers. Another important point of a shopping activity is the preparation of a shopping list, which helps the user to remember the things that need to be bought. We will present an implemented web-based agenda, where users or user groups can enter tasks, such as buying certain articles for a party. Typical existing organizers, such as PDAs or smartphones, only provide an alarm function to remind one user at a certain time. Our novel "Ubiquitous Agenda Service" allows the user to specify a place, or select a semantic category, such as a supermarket or brand. As soon as the user is nearby one of the specified places, a location-aware mobile device can present a reminder. Inside the shop, public displays will recognize the mobile device wirelessly via bluetooth and adapt their advertisement to the user's shopping list and general interests. In this chapter, we will put a focus on the role of cognitive and affective states for the adaptation of information presentation in instrumented environments. We will present how to recognize the resources of users with specialized "Dynamic Bayesian Networks" that probabilistically estimate the cognitive load and the time-pressure of users on the basis of their symptomatic behavior and physiological data that is derived from bio-sensors that measure for example the heart rate, the muscle tension, the electrodermal activity, or the eye movements. Finally, we will present an ontological approach to model and share the limited cognitive resources of users between different resource-adaptive applications. The "Ubiquitous User Model Service" provides contextual information on the users' actions, characteristics, and locations, while the users are enabled to access and control their profiles via a sophisticated web interface which integrates the necessary privacy issues.

1.1 Overview Described Within a Motivating Scenario

In order to get an impression of what shopping might look like in the future, imagine a fictitious shopping scenario in an instrumented environment in which Mrs. Smith and her husband are consumers. In preparation for her shopping, Mrs. Smith creates an electronic shopping list using a web interface. At the entrance of the supermarket, Mrs. Smith connects to her current shopping list with the tablet PC mounted at the handle of her shopping cart. She is navigated through the supermarket by an indoor navigation system (see chapter *Seamless Resource-Adaptive Navigation*). Meanwhile, Mr. Smith remembers that he has invited a friend for dinner and recognizes that he has no more wine at home. Thus he adds the entry "some French wine" to his wife's electronic shopping list which immediately appears on the screen of her shopping cart. When she sees the new entry, Mrs. Smith heads for the wine department to which she is guided by a projected virtual character that moves along the shelves and walls of the supermarket (see Sect. 6). When she enters the wine department, Mrs. Smith notices an elderly lady standing in front of the interactive wine information kiosk asking for some wine that suits her taste. On the basis of her speech input, the

kiosk system recognizes that the questioner is an elderly woman and recommends preferably sweet wines in a slow and comfortable voice (see Sect. 7). As Mrs. Smith does not have much knowledge of wine (this information can be retrieved from an ubiquitous user model, see Sect. 8) and the kiosk is already occupied, she uses the Mobile ShopAssist on her PDA to get some information about the different wines she considers buying (see Sect. 3). The Mobile ShopAssist monitors the user's choice and matches her actions to an affective state model (see Sect. 9). Beside other interaction modalities for the system output, a wine bottle that Mrs. Smith takes out of the shelf can "answer" her questions explaining her its features (see Sect. 2). In this way, Mrs. Smith can learn more about the specific features of the wines and compare them with each other. The required information can also be presented visually on projected displays that automatically appear at the back surface in the shelf when a bottle is taken out of it (see Sect. 4). If Mrs. Smith is still undecided which wine to buy after some time, an ambient sound notification system seamlessly informs an employee of the shop who is a wine expert that there is a customer in the wine department who might need some help (see Sect. 5).

2 Dialogue Shell of Talking Products

One important design goal of our interactive shopping assistance is to support arbitrary users, particularly computer novices, who are not able or willing to learn the use of such a system. We therefore have to find a solution that provides a natural interaction, requiring minimal effort of a user to understand and utilize the assistance system. Nijholt et al. [43] suggest that a limited animistic design metaphor seems to be appropriate for human–environment interaction with thousands of networked smart objects. People often tend to treat objects similar to humans, according to findings of Reeves and Nass [50], which allows users to explain the behavior of a system if they lack a good functional conceptual model. In consequence, we decided to employ a natural language system, which enables the user to talk to each product.

Our group conducted a usability study of a multi-modal shopping assistant [62]. The implemented system allows users for instance to request product information in a combination of speech and selecting gestures (i.e., taking a product out of the shelf). Findings of this study showed among others that users generally preferred direct over indirect interaction, i.e., by asking "What is your price?" instead of "What is the price of this camera?" which encouraged us to pursue this approach.

Previous studies have shown that interacting with embodied conversational agents that have consistent personalities is not only more fun but also lets users perceive such agents as more useful than agents without (consistent) personalities [42, 20]. It is further shown that the speech of a consistent personality enables the listener to memorize spoken contents easier and moreover reduces the overall cognitive load [23, 42]. Thus we emphasized the anthropomorphic aspect of this interaction pattern by assigning personalites to products, which are reflected by the spoken responses of a product.

Product manufacturers benefit as well, since the personalization of the product provides a new channel to communicate a brand image or distinct attributes of a certain product. A study within the context of marketing research showed that if in radio advertisements a voice fits the product, it helps the listener to remember the brand, the product and claims for that product [44].

2.1 Modelling Personality in Voices

In a first step we created voices that reflect certain personalities according to Aaker's brand personality model [1] only by adjusting prosodic parameters. We chose this model over the (rather similar) five factor model [37] commonly used in psychology, since we are applying the concept of talking objects in the shopping domain. However, both models are rather similar and to a certain extent exchangable.

We changed the four prosodic parameters pitch range (in semitones), pitch level (in Hz), tempo (as a durational factor in ms), and intensity (soft, modal, or loud as in [57]) according to our literature review [55]. For example, a competent voice has a higher pitch range (8 semitones), a lower pitch level (–30%), a 30% higher tempo, and a loud intesity compared to the baseline voice. In [55] we also evaluated whether it is possible to model different personalities with the same voice by adjusting these prosodic parameters, such that listeners will recognize the intended personality dimension. The study has shown that there are clear preferences for our prosody-modeled speech synthesis for certain brand personality dimensions. But not all personality dimensions were perfectly perceived as intended, such that we have to amplify the effect.

Personality is certainly not only expressed in qualitative attributes of a voice, other properties of a speech dialogue are also essential, like the used vocabulary or the general discussion behavior. For this reason we created a dialogue shell that incorporates these aspects.

2.2 Expressing Personality in Dialogues

The widely adopted personality model by Costa and McRae [37] constitutes five dimensions of human personality: Extraversion, Agreeableness, Conscientousness, Neuroticism, and Openness on a scale from 0 to 100. Obviously, differentiating 100 levels in a dimension is far too much for our goals, therefore we simplified this model by discriminating three levels in each dimension:

- low: value between 1 and 44 (31% of population)
- average: values between 45 and 55 (38% of population)
- high: values between 56 and 100 (31% of population)

Related work, e.g., by Andre et al. [2] limited their personality modeling to only two of the five dimensions, namely extraversion and agreeableness, since these are

the most important factors in interpersonal communication. Nevertheless, we discovered considerable influences of openness and conscientousness to speech, therefore we incorporated these two dimensions as well. The effect of the dimension neuroticism is mainly to describe the level of susceptibility to strong emotions, both positive and negative ones [17]. It is further shown that the level of neuroticism is very hard to determine in an observed person [26]; thus we decided that four dimensions will suffice for our work.

We conducted an exhaustive literature review on how speech reveals different personality characteristics. Among numerous other resources, two recent research papers provided essential contributions to our work: Pennebaker and King's analysis in *Journals of Personality and Social Psychology* [48] and Nowson's *The Language of Weblogs: A Study of Genre and Individual Differences* [45]. In both studies a large number of text blocks were examined with an application called Linguistic Inquiry and Word Count[1] (LIWC), which analyzes text passages word by word, comparing them with an internal dictionary. This dictionary is divided into 70 hierarchical dimensions, including grammatical categories (e.g., noun, verb) or affective and emotional processes. Pennebaker determined in a study the 15 most reliable dimensions and searched for them in diary entries of test persons with LIWC. With these results together with the given personality profiles of the probands (according to the five factor model), he identified correlations between the two. Nowson performed a similar study and searched through weblogs for the same LIWC factors.

Based on these results, we provided a set of recommendations on how responses of a talking object with a given personality should be phrased. For instance, for a high level of extraversion these recommendations are given:

- Prefered bigrams: *a bit, a couple, other than, able to, want to, looking forward*, and similar ones.
- Frequent use of terms from a social context or describing positive emotions
- Avoidance of *maybe, perhaps*, and extensive usage of numbers
- Usage of colloquial phrases, based on verbs, adverbs, and pronouns
- Comparably more elaborate replies

Following these principles we implemented basic product responses (greetings, inquiries for product attributes, farewell) for several personalities. All possible replies of our dialogue shell are stored in one XML-file, which we named the *Anthropomorphic Fundamental Base Grammar*. All entries include an associated personality profile, for example:

```
<reply
    query="hello"
    reply="Hello, nice to meet you!"
    ag="1" co="2" ex="1" op="1">
<\reply>
```

[1] http://www.liwc.net/

which means that this is the greeting of a product with average agreeableness, extraversion, and openness and a high value in conscientousness. Another example:

```
<reply
    query="hello"
    reply="Hi! I'm sure I can help you! Just tell me
        what you need and I bet we can figure
        something out!"
    ag="2" co="2" ex="2" op="2">
<\reply>
```

All entries that do not regard any particular personality should have average personality values in all dimensions.

A central product database with all products and their attributes is extended by the assigned personality profile, i.e., the values in each of the four dimensions. When the application starts up, the dialogue shell retrieves the product data of each product instance and extracts the appropriate entries from the base grammar to build the custom product grammar. If there are no entries that exactly match the given profile, the one that has the most identical values will be chosen. This dialogue shell generates a consistent speech interface to a product by knowing its attributes and a given personality profile, for instance preset by the manufacturer.

3 Mobile ShopAssist

The Mobile ShopAssist (MSA) is a platform originally designed to demonstrate a wide range of different multimodal interaction possibilities in everyday contexts, and particularly those contexts in which a user is mobile [61]. Created for use in mobile and ubiquitous environments, the ShopAssist application allows shoppers, accompanied by a PDA, to enquire about product features and to compare different products with one another. This is achieved through the use of input modalities like speech, handwriting, and selection gestures. Figure 1 shows the ShopAssist application in use during field studies conducted at Conrad Electronic in Saarbrücken.

Since its conception, the MSA has become a test bed for a number of research focuses including mobile multimodal interaction, on- and off-device input recognition, on- and off-device presentation output planning, anthropomorphisation, and public associated displays. The architecture of the platform, as implemented for demonstration of mobile multimodal interaction, can be seen in Fig. 2.

Multimodal interaction refers to "the means for a user to interact with an application using more than one mode of interaction" [60]. Such interaction might occur sequentially or simultaneously in time, and may also contain semantically overlapped information in which certain semantic constituents (such as a shopping product's price) is provided multiple times by similar or different modalities (such as speech and handwriting).

Fig. 1 Mobile ShopAssist interaction, as used during field studies at Contrad Electronic in Saabrücken

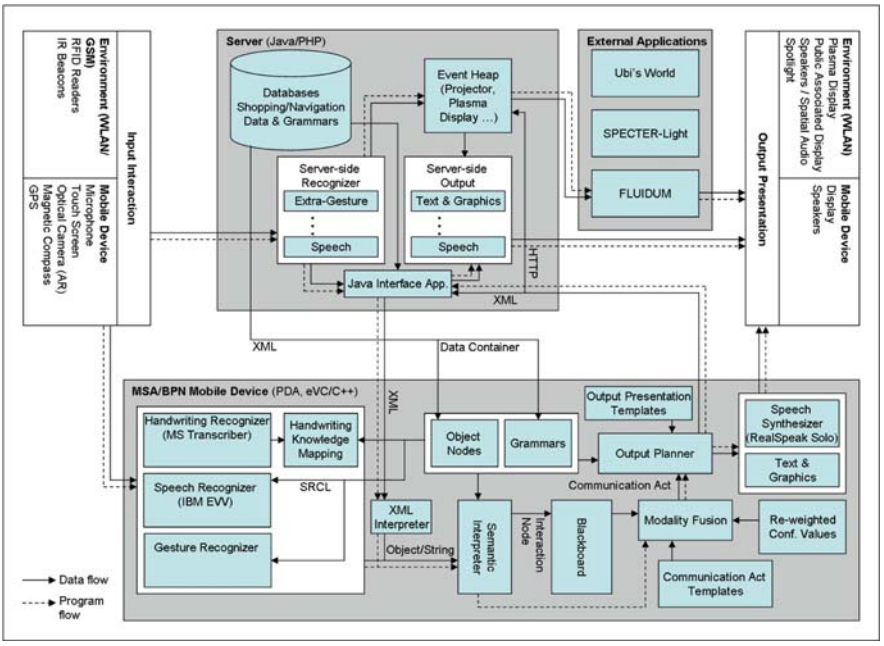


Fig. 2 The MSA architecture showing the data flow between different components

The MSA supports an Always Best Connected (ABC) methodology such that interaction components located in a publicly instrumented environment, for example, distributed speech recognizers and gesture recognizers, can be made accessible to a user for the purpose of enhanced application functionality like improved recognition accuracy and support for larger vocabularies. This adaptation to available resources (i.e., recognition results from multiple recognizers) has been made possible through results from field studies that were conducted to determine correlations

between recognizer confidence and recognition accuracy. The benefit of such correlations is that they allow for recognizers that inherently process signals differently (e.g., speech and handwriting) to have their results compared with one another. This also applies to same-type recognizers (e.g., server-side and embedded speech engines) in which confidence values are generally based on entirely different factors like acoustic models, grammars, and associated computing power.

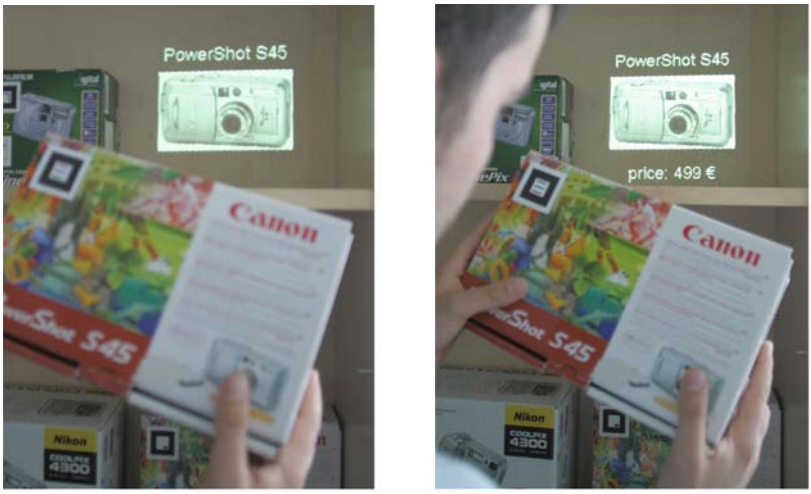
Similar to the recognition of user input, output presentation planning in the MSA is also resource-adaptive. In particular, user parameters from UbisWorld (e.g., age, gender, and modality preference) as well as context parameters (e.g., spoken dialogue tempo, SNR, ambient light level, and number of surrounding people) are used to determine particular system reactions. Typical reactions of the system are for example the display duration of textual output, the format and tempo of speech output, and whether output is to be presented on-device or off-device (i.e., on the mobile PDA, or on devices located in the surrounding instrumented environment).

4 Product Associated Displays

PADs (Product Associated Displays) are projected virtual displays which are created at locations that can intuitively be associated with the products the user is currently interacting with. Instead of displaying product information on a stationary screen, which can be installed, e.g., beside the product shelf, and which the user might not be aware of because of the spatial distance to the product, a projected PAD presents the relevant product information in the gap left in the shelf when the product is taken out of it. As in the process of taking a product, its former location in the shelf remains in the user's peripheral view, a PAD that occurs there immediately after the action is very likely to catch the user's attention. In this way, a spatial mapping between the physical location and the displayed information is established and a relationship between the product and the corresponding information on the PAD arises automatically.

In our shopping scenario, PADs are projected using the Fluid Beam system [59]. Its hardware part consists of an LCD projector and a digital camera placed in a moving yoke in which they can be rotated horizontally (pan) and vertically (tilt). In this way, the projector beam can be directed at almost any surface in the room. In order to avoid image distortion due to oblique projection, the Fluid Beam software implements a method described in [49]. It is based on the fact that projection is a geometrical inversion of the process of taking a picture given that the camera and the projector have the same optical parameters and the same position and orientation. The implementation of this approach requires an exact 3D model of the environment, in which the projector is replaced by a virtual camera. By synchronizing the movements of the steerable projector in the physical environment and the virtual camera in the 3D model, the image delivered by the virtual camera appears undistorted when it is projected in the physical environment. Thus virtual displays showing images, videos, or video streams can be placed in the 3D model and they

are projected at the corresponding locations in the physical world when the virtual camera (and respectively the steerable projector) is directed at them. In this way, a sort of virtual layer is created that covers the surfaces of the physical environment, on which projected virtual displays can be placed and moved.



(a) Initial display showing a picture of the product and its name

(b) Price information displayed on a PAD as an answer to the user's request

Fig. 3 Product associated display

The event of taking a product out of a shelf or putting it back is recognized by means of passive RFID tags attached to the products and an RFID antenna placed behind the shelf. If the user takes out a product, this action is recognized by the system and a corresponding event is generated and sent to an Event Heap [29]. The Mobile ShopAssist [63] receives the event from the heap and sends a command to the steerable projector to display a PAD at the appropriate location showing a picture of the removed product and its name (see Fig. 3a). After that, the user is given the opportunity to ask for additional information about the product (e.g., price) using speech, handwriting, or intra-gestures on his or her PDA (see Sect. 3). The answer to the user's request can then be displayed on the PAD if this is allowed by the user's preference settings (see Fig. 3b).

5 Personalized Ambient Soundscape Notification

In most instrumented environments the visual sense is the primarily used of all human senses. Usually, audio signals are limited to simple warning cues and system feedbacks that are in most cases intrusive because of their dissimilarity compared to the environmental noise. That has the effect that persons present in the room will be distracted from their current tasks. To prevent the disturbing effect of traditional

notification signals we developed the novel concept of non-speech audio notification embedded in ambient soundscapes to provide a method for multi-user notification in a more discreet and non-disturbing way.

5.1 Introduction to Ambient Audio Notification

In 1978, English musician Brian Eno coined the term ambience in combination with music in the notes to his longplayer *Ambient 1: Music For Airports*. This type of music has a calming effect and can be listened to either actively, that means the focus of attention lies on the music, or it can be listened to peripherally without paying attention to the music. This effect is also known as the auditive figure-ground phenomenon which decribes human's ability to pay attention to an auditory stream (figure) while at the same time any other sound is listened to peripherally (ground) [10]. Already in the year 1953, Colin Cherry described this effect in his famous "cocktail party" experiment when he found out that the auditive perception is associated with the attention of a person [16]. The allocation of the limited resource attention depends on a variety of factors like the stimuli that act on the person and his current mental and social conditions [21].

Perception of auditive signals can be divided into the physiological phenomenon of hearing and the semantic sound processing which leads to the personal interpretation of the signal, influenced by the experiences of each individual listener. The intensity and complexity of environmental noises influence whether we perceive a single sound or whether it is masked which depends on multiple factors like loudness and the frequency of the noises. Traditional audio notification signals are mostly stand-alone cues that attract the attention of everybody in a room because they are not integrated into the natural sound environment [46]. That works fine for high-priority notifications (e.g., fire alarm), but often a more personal and discreet notification is desirable.

We had two main goals for the design of our notification signals. On the one hand we want to seamlessly integrate the notification signal into background music without arousing the attention of other people, but on the other hand the target person must become aware of the signal.

Auditory experiences can be permanently extended and trained [3]. We use this fact to make the listener more sensible to his specific auditory signals that we use for attracting his attention. These audio cues are used to provide the listener with information that he links with the specific auditory signal. The user can choose which sound he wants to link with which information, so we get an individual and personalized notification that respects the user's preferences.

Since only the user knows which sound he selected for which information, this type of notification also slightly fulfills the privacy aspect.

5.2 Ambient Soundscapes and Audio Notification Cues

The main problem with traditional stand-alone notification signals is the distraction of other present persons, especially in multi-user environments. Indeed, popular

non-speech audio cues like earcons [6] and auditory icons [11, 14] can provide a perceptible type of notification but they are also separated from environmental noise.

To introduce more privacy and confidentiality, we decided to integrate notification instruments with respect to the musical compositions seamlessly into background music, the *ambient soundscape*, which serves as the musical envelope [13]. Instead of attracting the listener's attention, the soundscape should have a calming and mood influencing effect (see also [4, 35]).

To reach this goal we composed and recorded three ambient soundscapes and suitable notification instruments by ourselves. We took some perceptual constraints such as the auditive Gestalt laws and several studies dealing with musical perception into consideration as described in [15, 51, 52, 19]. Table 1 gives a brief overview of the emotional impact of compositional parametes on the listener's mood. In our shopping scenario, the ambient soundscapes create a more friendly atmosphere for consumers.

Table 1 Categorization of musical parameters, including range and emotional impact [12]

Category	Parameter	Range	Emotional impact
Time	Speed	fast – slow	pleasant – calm
	Phrasing	staccato – legato	lively – gently
	Rhythm	firm – smooth	serious – dreamy
	Dynamic	cresc. – decresc.	animated – relax
	Meter	even – odd	dignified – restless
Pitch	Mode	major – minor	bright – plaintive
	Frequency	high – low	exciting – sad
	Melody	ascending – descending	dignified – serene
	Note Range	$\geq$octave – $\leq$octave	brilliant – mournful
	Harmony	consonant – dissonant	serene – ominous
Texture	Volume	forte – piano	animated – delicate
	Orchestration	instrumentation	majestic – grotesque

In the second phase we add *notification instruments* to the ambient basic soundscape and play them with slightly increased volume at the current position of the task person by using an indoor positioning system [58] and a spatial audio framework [54]. The *Always Best Positioned* (ABP) mobile localization system called *LORIOT* uses RFID technology in combination with infrared beacons to find out what the user's current position is. The calculation is done on the PDA by using Dynamic Bayesian Networks (DBN's) [8]. More information about the positioning system can be found in the chapter "Seamless Resource-Adaptive Navigation" of this book. *SAFIR* (Spatial Audio Framework for Instrumented Rooms) is used to play the audio cues at the loudspeaker that is the nearest to the target person's postition.

Since the notification instruments will be seamlessly integrated in the ambient soundscape this has the effect that an occurring notification could be perceived after a while. To prevent the effect of ignoring a notification, we also provide a hierarchy of notification signals that are grouped by "level of intrusiveness" [33], depending on the importance of the occurring event.

1. High-Priority
   Signals: Arousing Noises (e.g., beep, siren, and bell).
   Immediate and intrusive notification which is independent of the current compositional context.
2. Medium-Priority
   Signals: Ambient Noises (e.g., birds, rain, water- and wind noises).
   Immediate and context independent, but still an ambient notification with natural sounds.
3. Low-Priority
   Signals: Notification Instruments (e.g., drums, cymbals, guitar, piano, and violin).
   Seamless integration of melodic patterns, played by natural instruments, into the ambient soundscape (compositional-context-awareness).

The user can choose a soundscape that matches his personal preferred music style and an instrument or ambient noise that he can easily recognize. His personal preferences can be stored in his user profile of UbisWorld, a user model ontology [28] that is also described in this chapter. The profile information can be accessed by the notification system via http request.

The effectiveness of the peripheral perception was successfully tested in a user study with 25 persons [30] where we especially checked whether the users percept the notification instruments and the elapsed time to recognize the notification (delay time). The study was subdivided into a computer-based test and a questionnaire to get a subjective and personal feedback of the participants' opinion about the soundscapes and this new concept of notification.

5.3 Applications and Shopping Scenario

The *Personal Ambient Audio Notification* service (PAAN) handles an audio server, the data exchange to UbisWorld, the indoor positioning system (LORIOT), and the spatial audio system (SAFIR). Figure 4 shows the hardware that we use for PAAN. The audio hardware includes Hi—Fi amplifiers and loudspeakers that are connected to a multi-channel soundcard. For a scenario where the user changes his position, we use the Always Best Positioned system LORIOT that uses a PDA equipped with wireless LAN and an RFID reader card, active RFID tags that are mounted on the ceiling of the room and optional infrared beacons mounted at shelves or walls.

In our shopping scenario, the ambient soundscape can be selected by an employee of the wine store by browsing available soundscapes in the web interface on his computer. The soundscapes are stored on an audio server and managed by an audio database. Search queries for the audio database can include the name of the soundscape or *GEKOS*[2]-keywords that describe each soundscape by its compositional elements. Employees of the store can change their personal audio notification

[2] GEKOS: **G**enre, **E**xpression, **K**ey, **O**rchestration, **S**ignature

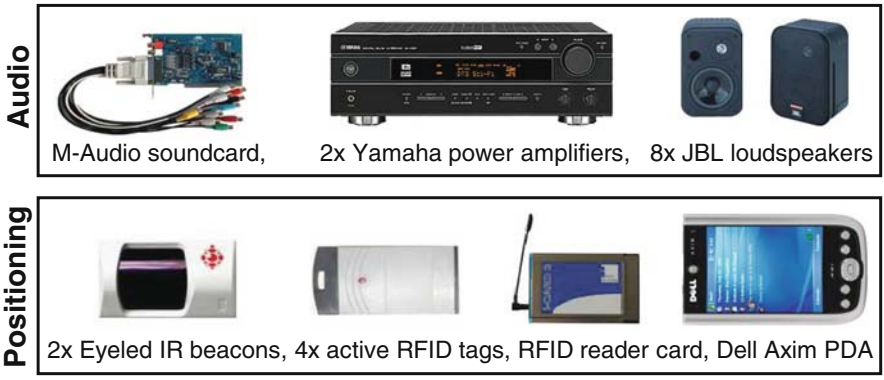


Fig. 4 Audio- and positioning hardware requirements for PAAN

instrument by updating their UbisWorld account. The IP address of the user's PDA can also be specified and exported in an XML file which can be accessed by PAAN via http request to route an event notification to the appropriate task person [31].

Figure 5 shows an audio sequence of a possible shopping scenario with two selected Notification Instruments (NI 1, NI 2) assigned to two employees, an Ambient Noise (AmN) for group notification and an Arousing Noise (ArN) for high-priority notification. The notification service reacts to relevant events (E_i) by mixing the adequate notification in the playing soundscape at the right time [32].

The Ambient Soundscape (AS) starts automatically when a registered user, namely an employee, enters the instrumented area of the shop with his PDA ($E_0(t_1)$).

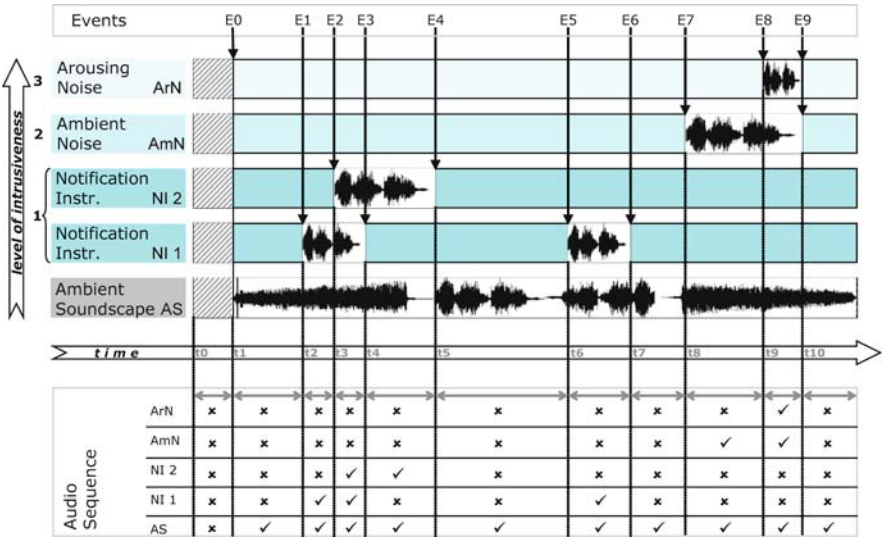


Fig. 5 Audio sequence example

The preselected notification signals that are assigned to authorized users are in stand-by mode (muted).

Assume a consumer who enters the wine department of a supermarket with the intention of purchasing a French red wine while at the same time the employees are out of sight and do not notice the appearance of the potential consumer. After his arrival, the customer ($E_1(t_2)$) can be detected for example by the instrumented shopping cart [56] or a location-aware PDA [58]. The notification system determines the salesman's current position by checking the positioning coordinates of his PDA, matches them to the nearest loudspeaker and informs him about the presence of a person by starting to play his personal notification instrument (NI 1) with slightly increased volume at his position. The salesman notices that his instrument (e.g., piano) starts playing in the soundscape and that this is the appointed signal for a potential consumer in the wine department. Back in the wine area, it turns out that he cannot satisfy the consumer's wish because he is looking for a specific red wine and needs the advice of a wine specialist focused on French red wines. Thereupon, the salesman calls the specialist by starting NI 2 (e.g., drums), which is the signal for the French wine specialist to come to the wine department ($E_2(t_3)$). After whose arrival, the salesman and the specialist stop their notification instruments by pressing a GUI button on their PDAs ($E_3(t_4)$, $E_4(t_5)$). The instruments leave the music seamlessly and only the basic soundscape is still playing. Event $E_5(t_6)$ describes a similar scenario where the salesman receives a call to come to the department where he stops the notification after his work is done.

Event $E_7(t_8)$ is an example for group notification and could occur if the head of the wine department wants to call his colleagues for a meeting by sending an email to the department's mail account. The Ambient E-Mail Notification system (*AEMN*) periodically checks the mail server for incoming messages and filters them for predefined keywords. The important announcement email triggers an event with the corresponding group notification signal in the form of an ambient water noise (AmN) which immediately starts playing in the whole department. Unfortunately, not every staff member noticed the ambient notification after a while, so AEMN sets the level of intrusiveness to the highest level and starts playing an additional Arousing Noise (ArN), e.g., a beep sound ($E_8(t_9)$). After the arrival of the remaining employees, the two notification sounds were stopped on the PDA or on the department's desktop PC at time t_{10}.

The introduced personalized ambient notification is an effective and non-intrusive concept to provide users with information location-aware and under low-privacy aspects.

Now, music is no longer a pure emotion mediator, but rather contain emotion and content, whereby the sum of these two factors results in the information content of the music.

6 Virtual Room Inhabitant

The Virtual Room Inhabitant (VRI) is a virtual character capable of guiding and following a user throughout physical spaces. The main purpose of the VRI in our shopping scenario is to welcome the user when entering the shopping area and

to guide the user to a particular shelf within the room. Unlike traditional virtual characters, the VRI is not limited to the narrow boundaries of a display or a fixed projection. Instead, the VRI is free to move along arbitrary surfaces in its surroundings and to appear at any location that will allow for a projection. From a technical point of view, the VRI utilizes a steerable projector (as described in Sect. 4) and a spatial audio system (see Sect. 2). This combination allows to visually locate the VRI within physical spaces and to locate the origin of the characters voice at the same location, hence conveying the impression of the character actually standing at that location.

In order to allow the VRI to follow the user throughout the room, it is necessary to sense the users location and to react to the users movements accordingly. The position is acquired using a positioning technology based on PDA localization. The position is determined by using a combined system of active RFID tags and infrared beacons. Since the infrared beacon technology demands a direct line of sight between sender and receiver, we also get a rough estimate of the users orientation (the details of the indoor positioning system are described in [9]). The positioning information is stored on the so-called Event Heap which implements a blackboard architecture and allows clients to post and retrieve data by signing in to particular information channels. The character engine which constitutes the central part of the VRI registers itself to the positioning information channel on the Event Heap. If the user enters a particular region in the shop, the situation is recognized by the character engine. In order to start the VRI, the character needs to get access to the necessary hardware. Therefore, it posts a request to the presentation manager which, in combination with a device manager, grants access to all registered devices (see Fig. 6 for details). Each device in our hardware set up has to register itself and services it offers at the device manager. The presentation manager handles all device requests from concurrently running applications in the environment. The presentation manager decides, whether a particular user may access a device or

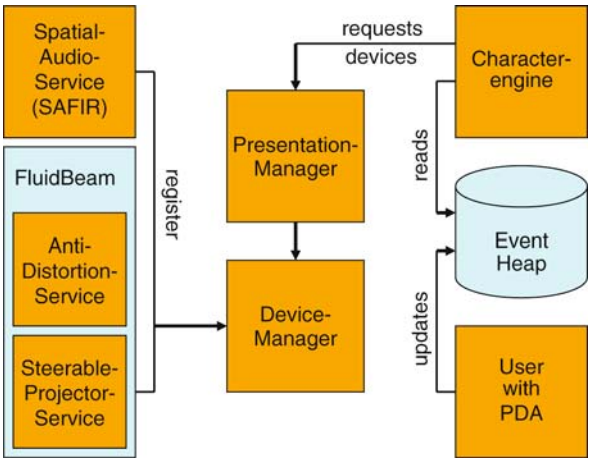


Fig. 6 The system components of the Virtual Room Inhabitant

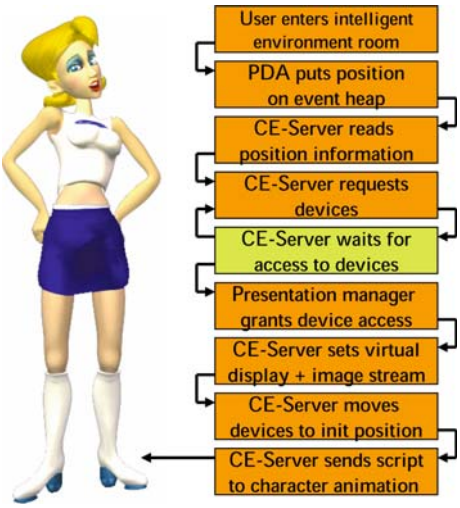


Fig. 7 The initialization sequence for the Virtual Room Inhabitant

whether the request is denied or delayed. The remote access mechanism is realized with Java Remote Method Invocation[3] objects, which allow arbitrary applications to control remote devices as if they were locally connected. The whole device access mechanism is depicted in Fig. 7. The character engine consists of two parts, namely the character engine server (CE-server) written in Java and the character animation, which was realized with Macromedia Flash MX.[4] These two components are connected via an XML-Socket-Connection. The CE-server controls the Flash animation by sending XML commands/scripts. The Flash animation also uses the XML-Socket-Connection to send updates on the current state of the animation to the CE-server (i.e., whenever a part of an animation is started/finished). The character animation itself consists of ∼9,000 still images rendered with Discreet 3D Studio Max[5] which were transformed into Flash animations. To cope with the immense use of system memory, while running such a huge Flash animation, we divided the animation into 17 subparts. While the first consists of default and idle animations, the remaining sixteen are combinations of character gestures, like for example, shake, nod, and look behind. Each animation includes a lip movement loop, so that we are able to let the character talk in almost any position or while performing an arbitrary gesture. We have a toplevel movie to control these movie parts. Initially, we load the default movie (i.e., when we start the character engine). Whenever we have a demand for a certain gesture (or a sequence of gestures), the CE-server sends the corresponding XML script to the toplevel Flash movie which than sequentially

[3] http://java.sun.com/products/jdk/rmi/
[4] http://www.macromedia.com/software/flash/
[5] http://www4.discreet.com/3dsmax/

loads the corresponding gesture movies. The following is a short example of an XML script for the character engine:

```
<VRI-script>
    <script>
        <part>gesture=LookFrontal sound=welcome1.mp3</part>
        <part>gesture=Hips sound=welcome2.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
    <script>
        <part>gesture=LookFrontal sound=cart.mp3</part>
        <part>gesture=PointDownLeft sound=cart2.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
    <script>
        <part>gesture=PointLeft sound=panel.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
</VRI-script>
```

Each script part is enclosed by a script tag. After a script part was successfully performed by the VRI animation, the CE-server initiates the next step (i.e., move the character to another physical location by moving the steerable projector and repositioning the voice of the character on the spatial audio system, or instruct the VRI animation to perform the next presentation part). In order to guarantee a smooth character animation, we defined certain frames in the default animation as possible exit points. On these frames, the character is in exactly the same position as on each initial frame of the gesture animations. Each gesture animation also has an exit frame. As soon as this frame is reached, we unload the gesture animation, to free the memory, and instead continue with the default movie or we load another gesture movie, depending on the active XML script.

In addition to its animation control function, the CE-server also controls the spatial audio device, the steerable projector, and the antidistortion software. The two devices, together with the antidistortion software are synchronized by commands generated by the CE-server, in order to allow the character to appear at any position along the walls of the room, and to allow the origin of the character's voice to be exactly where the character's visual representation is.

Presentations are triggered by the user's movements. As soon as a user enters the instrumented room, the CE-server recognizes the relevant information on the Event Heap. On the next step, the CE-server requests access to the devices needed for the VRI. Given access to these devices is granted by the presentation manager (otherwise the server repeats the request after a while), the CE-server generates a virtual display on the antidistortion software and starts a screen capture stream, capturing the character animation, which is then mapped on the virtual display. It also moves the steerable projector and the spatial audio source to an initial position.

As a final step, the CE-server sends an XML script to the character animation, which will result in a combination of several gestures, performed by the character

Fig. 8 The Virtual Room Inhabitant in action

while playing synchronized mp3 files (synthesized speech) over the spatial audio device. The whole initialization process is indicated in Fig. 7.

The main purpose of the VRI is to guide the user while exploring a shop; however, it may also perform references to physical objects, for example, to help the user to locate a particular product. In the example in Fig. 8, the VRI is standing between a wall-mounted display and a product shelf. At this location it may point both at objects on the screen as well as on products in the shelf.

7 Live Acquisition of User Profile Data from Speech

As of today, the most common case in modeling shop visitors is that no predefined information is available about the individual subject, possibly because it is their first visit to that specific shop, or it could be that they did not spend the time creating a profile yet. But even if the visitor does have a user profile provided to the shop by a compatible service such as *UbisWorld* (see Sect. 8), he may not be immediately authenticated to the system when he enters the shop, be it for privacy reasons or just because he just does not see the necessity yet. It resembles a behavior that is well known from the web, where users often do not log into websites unless a privileged action requires them to do so.

There are a number of advantages with having more information about the visitor. First and foremost, more profile data allow for a more personalized experience. However, requiring the user to enter some minimal information manually, which is certainly an option, carries the risk of lowering the overall customer satisfaction, as having to fill out a form (e.g., on a mobile device) is at least time-consuming. For some types of visitors such as elderly people, it may be even more inconvenient and more likely a reason to avoid such a shop or completely refrain from using its digital services.

Instead, in this case where no previous knowledge about the visitor is present, it is desirable that all available sources of user characteristics that do not require direct interaction of the person be utilized to draw as many conclusions as possible in order to bootstrap a user profile. Speech is one such source. For this purpose, a set of speech-based classifiers have been developed in the AGENDER project, which can classify the age, gender, and language of a speaker with relatively high accuracy.

A person's recorded voice contains a lot of information about the speaker. Literature studies conducted by Müller [40, p. 43] suggested that voices do not only differ between the genders and between different ages, but that also a gender-specific vocal aging can be witnessed. The information is conveyed in prosodic features such as *pitch*, *jitter*, *intensity*, *shimmer*, and *harmonics-to-noise ratio*. Using methods from signal processing implemented in the tool *Praat*,[6] common statistics based on these features (e.g., mean and standard deviation) were extracted on large corpora of labeled speakers such as *Timit* [24] and *BAS*[7] [53]. A Gaussian probability distribution analysis performed on the extracted data revealed the subgroups of features that were most promising in being a discriminator for certain classes. In subsequent tests, it also became apparent that some classes, including those spanning different speaker properties, could be grouped to form a single combined class that resulted in a better overall performance for a specific feature set. For example, one classifier discriminates between the three classes "children," "female adults," and "male adults or seniors."

Based on these features, classifiers for each of the combined classes were trained using different machine learning algorithms. Most of them were from the *WEKA*[8] machine learning toolset. The current implementation uses a fast Gaussian Mixture Model for classification. The results from multiple classifiers extracted on a "first layer" can be combined on a "second layer" using a Dynamic Bayesian Network such as the one depicted in Fig. 9 [7]. This method can also be used to exploit the aforementioned fact of gender-dependant vocal aging by modeling the probability of a gender-specific age classifier as dependant on the probability output of the gender classifier. This improves performance because currently, gender can be classified with a much higher accuracy than age on unfiltered data. Additionally, the aspect of time is incorporated into the network when multiple utterances of the same speaker

[6] `http://www.praat.org`
[7] Bayerisches Archiv für Sprachsignale
[8] Waikato Environment for Knowledge Analysis

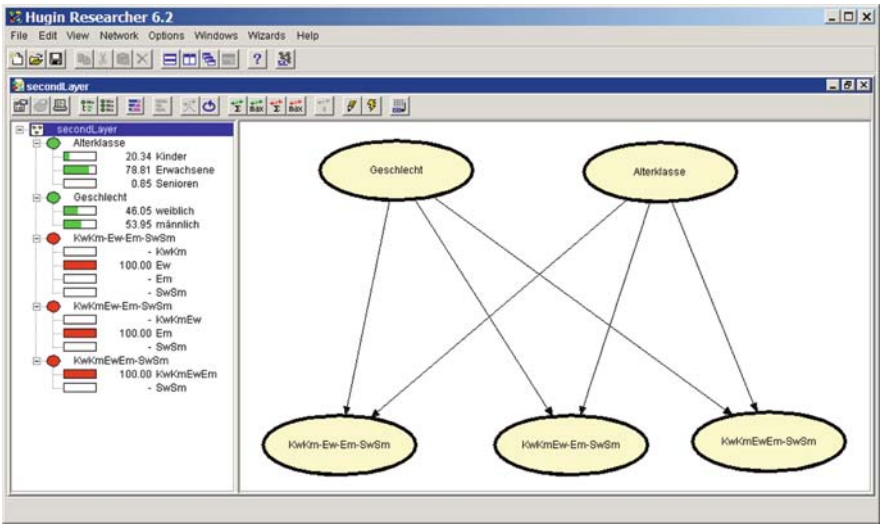


Fig. 9 A Bayesian Network which is part of the second layer in AGENDER is being evaluated with the application *Hugin Researcher*

are considered, so that the final probability should converge and reduce classification errors.

While there are no technical limitations relating to the number of age classes, only classification modules with two, three, and four age classes have been evaluated until now, with the four-class-classifier discriminating children (0–13 years), teenagers (14–19 years), young adults (20–64), and seniors (65+). One reason for this choice lies in the vocal significance of the chosen age borders. Another reason is the fact that with an increasing number of classes, it is considerably harder to come by an adequate amount of training material for each class. Table 2 shows the covariance matrix for an eight-class-scenario (combining age and gender, e.g., *Cf* = *Children female*). The average accuracy in that case is 63.5%.

For language, a different approach had been taken because prosodic features do not convey sufficient information for accurate language classification. The idea of a

Table 2 Confusion matrix for the 8-class-problem with an ANN. The total accuracy is 63.5% with a chance level of 12.5%. The diagonal axis is given bold

	8-class-problem					total accuracy 63.5%		
	Cf	Cm	Yf	Ym	Af	Am	Sf	Sm
Cf	**76.09**	4.07	13.6	5.06	0.54	0.05	0.44	0.15
Cm	54.25	**12.37**	12.52	15.51	1.13	0.25	3.78	0.2
Yf	54.15	2.41	**27.44**	13.16	1.28	0.1	1.37	0.1
Ym	20.08	3.98	6.33	**59.25**	1.03	1.13	4.96	3.24
Af	0.25	0	0.2	0.54	**84.73**	3.44	6.92	3.93
Am	0	0	0	0.74	3.53	**87.87**	1.57	6.28
Sf	0.59	1.13	0.15	2.5	3.78	0.93	**77.07**	13.84
Sm	0	0.05	0	1.67	1.18	1.47	12.47	**83.16**

phonotactics model in combination with a pseudo-syllable model first sketched in [41] has since been developed and extended. A large corpus containing speech samples in all languages to be considered was used to form the background model for the language classification problem. Then, a variable number of MFCC-based vector quantization front-ends were trained with the *HTK*[9] toolset on the background model or parts of it. Each of these front-ends is trained with different parameters and/or different speaker classes and can output a sequence of subphonemes (thus works as a segmenter, similar to a phoneme recognizer). From the output of the front-ends, n-grams were computed with $n = 1$, 2, and 3. In order to reduce the size of the models, only the statistically most significant n-grams were used, e.g. only 20% of the trigrams, but 100% of the unigrams. The actual feature used in the classifier is the relative count of the individual n-grams. Through experiments, the best front-ends were selected. Using all data of the background model, a normalization model was computed, with rank normalization providing the best results. For each language, the Support Vector Machine *SVM-Light*[10] was trained with a fixed training set using the corresponding (normalized) feature vector. The distribution of classes can be chosen freely. In a last step, using a test set, each of the classifiers was evaluated separately with different decision thresholds, which modify the output score, to minimize the mean error. The best-performing decision threshold was applied to the result of the final classifier. Decision thresholds can also be manually adjusted to improve the overall performance in scenarios where classes are not equally distributed.

The classifiers are implemented in a high-performance C++ library, which can be trained and configured at design-time to include any number of classifiers for the required classes. Using a tool named *SBC Development Platform*, these classifiers are compiled into so-called "embedded classification modules," which consist of mostly binary code that represents the classifier. There is also the option for including post-processing layers such as a Bayesian Network in these modules. This approach has already been successfully applied for gender-dependant recognition of age and is described in [40, p. 181].

With these classification utilities at hand, the next step to an application scenario is to identify a source from which speech will be taken and a suitable setup for the classification engine. In the shopping scenario, this could be a voice-controlled mobile application on a PDA like the *ShopAssist* (see Sect. 3) that guides the user through the shop and provides interaction with virtual item listings and actual products in a shelf. The input features that are used to classify the user are the utterances which make up the voice commands given to the *ShopAssist*. Using the same speech samples for classification is straightforward as it requires no additional effort on the user's side. An alternative to the PDA would be a device built into the shopping basket or a stationary microphone installed at a shop shelf facilitating communication with "Talking Products" (see Sect. 2). Also, a conversation with the Virtual

[9] Hidden Markov Model Toolkit, `http://htk.eng.cam.ac.uk`
[10] `http://svmlight.joachims.org`

Room Inhabitant (see Sect. 6) character could be used. In another typical scenario for AGENDER, which is automatic call forwarding in a call-center [22, p. 133], the voice recorded in an initial speech prompt serves as the classification input. It should be stressed that in a concrete situation, the input can be used at any time and in any order, or it may not be present at all. It is also possible to use multiple input sources if available. While there is the possibility to run most parts of Agender on a mobile device, a client/server-based approach where there is a single server handling all classification requests for all users is favored in scenarios that support it, because it will usually result in lower classification times. In the shopping scenario, the shop could set up a server running AGENDER and stream voice data used to interact with the *ShopAssist* from the PDA over wireless LAN or Bluetooth to that server.

While it does not lie within the domain of AGENDER how the obtained information is used, one common application that is also referred to in the shopping scenario from the beginning is user adaptation. By creating or updating a user profile, smart services consulted by the same user may be adjusted to the user's characteristics. For example, the virtual character may be chosen from the same age group as the speaker and answer in the correct language. For elderly people, it may be a good idea to reduce the rate of speech for easier understanding, which applies to the "Talking Products" as well. Another common scenario is to adapt the choice of products suggested to the user in other applications or advertisements to match the user's demographics, or even exert influence on the path created by indoor navigation systems that are part of the shopping experience. Adapting applications should not rely on the classification to be as reliable as information entered by the user himself, and – depending on the way the input is acquired – should provide fallback solutions if some information is not present, e.g., because the user did not provide any speech yet.

8 Ubiquitous User Modeling with UbisWorld

In order to realize user modeling for intelligent environment and ubiquitous computing as indicated by this future shopping scenario, the concept of *ubiquitous user modeling* has been proposed in [27]. This concept contains a RDF-based general user model ontology GUMO and a context markup language UserML that lay the foundation for inter-operability using Semantic Web technology. GUMO and UserML enable decentralized systems to communicate over user models as well as situational and contextual factors. The idea is to spread the information among all adaptive systems, either with a mobile device or via ubiquitous networks. UserML statements can be arranged and stored in distributed repositories in XML, RDF, or SQL. Each mobile and stationary device has an own repository of situational statements, either local or global, dependent on the network accessability. A mobile device can perfectly be integrated via wireless lan or bluetooth into the intelligent environment, while a stationary device could be isolated without network access. The different applications or agents produce or use UserML statements to represent

the user model information. UserML forms the syntactic description in the knowledge exchange process. Each concept like the user model auxiliary "has Property" and the user model dimension timePressure points to a semantical definition of this concept which is either defined in the general user model ontology GUMO, the UbisWorld ontology, which is specialized for ubiquitous computing, or the general SUMO/MILO ontology. Figure 10 shows Basic User Dimensions in the GUMO ontology.

GUMO collects the user's dimensions that are modeled within user-adaptive systems like the user's age, the user's current position, the user's birthplace, or the user's gender. In the GUMO ontology, long-term user model dimensions are categorized as demographics. Ontologies provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and widely spread application systems. Since ontologies have been developed and investigated in artificial intelligence to facilitate knowledge sharing and reuse, they should form the central point of interest for the task of exchanging situation models. Figure 11 shows an example of an ubiquitous user model in the UbisWorld.

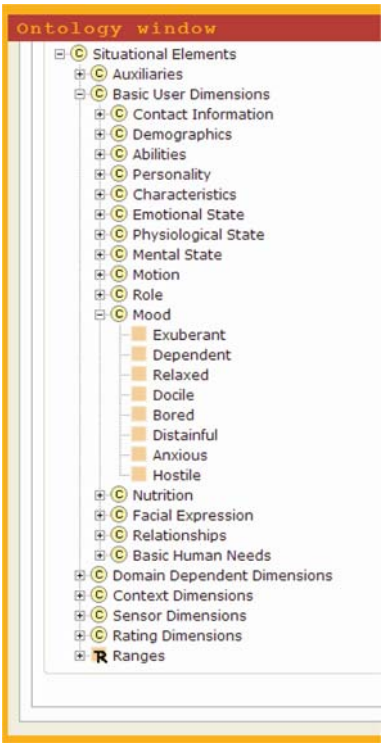


Fig. 10 Selected basic user dimensions in the GUMO ontology

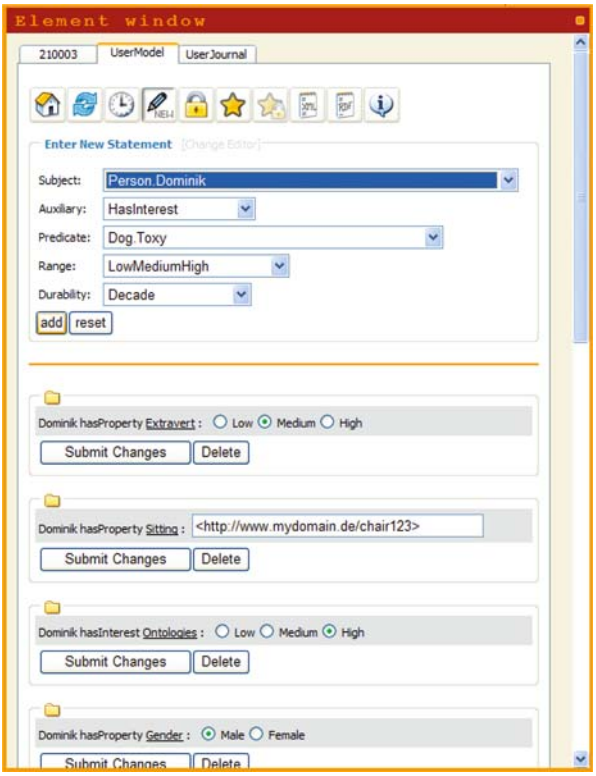


Fig. 11 User model inspection and editing in UbisWorld

The web ontology language OWL has more facilities for expressing semantics. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. Thus, OWL is our choice for the representation of user model and context dimension terms and their interrelationships. This ontology should be available for all user-adaptive and context-aware systems at the same time, which is perfectly possible via internet and wireless technology. The major advantage would be the simplification for exchanging information between different systems. The current problem of syntactical and structural differences between existing adaptive systems could be overcome with such a commonly accepted ontology. UbisWorld (Fig. 8) enables users to annotate their user models with the GUMO ontology. UbisWorld represents persons, objects, locations as well as times, events and their properties and features. UbisWorld could be understood as a virtual colored blocks world where each color represents a different category in the ontology. The main focus of this approach lays on research issues of ubiquitous computing and user modeling. Apart from the representational funtionality, UbisWorld can be used for simulation, inspection, and control of the real world.

9 Modeling Affect

The understanding of an users affective state surely enables new ways of how to adapt to a users specific situation. This is especially important in sales scenarios, where affect plays a major role. According to the current affective state of a person, she/he decides whether to accept a specific offer. By extending the underlying user model, respectively the ubis world ontology, by a fine grained model of affect more adapted sales dialog will be possible.

The representation and real-time simulation of affect appraises a users actions in the described scenario. As a result of this process possible short-term emotions and long-term moods will be computed.

9.1 Affect Taxonomy

The affect classes of the ubis world ontology is designed to represent and simulate affect types as they occur in human beings. As suggested by Krause [34] affect can be distinguished by the eliciting cause, the influence on behavior, and its temporal characteristics. Based on the temporal feature, we use the following taxonomy of affect:

(1) Emotions reflect short-term affect that decays after a short period of time. Emotions influence facial expressions, facial complexions (e.g., blush), and conversational gestures. (2) Moods reflect medium-term affect, which is generally not related with a concrete event, action, or object. Moods are longer lasting affective states, which have a great influence on humans' cognitive functions [39, 18]. (3) Personality reflects long-term affect and individual differences in mental characteristics [36].

As known by psychological research, those different types of affect naturally interact with each other. Personality usually has a strong impact on the emotions' intensities [5, 64]. The same applies to moods [18]. With our computational model we want to simulate the interaction of the different affect types in order to achieve a more consistent overall simulation of affect.

9.2 Affect Computation

Our work is based on the computational model of emotions (ALMA) described in [25]. It implements the model of emotions developed by the psychologists Ortony, Clores, and Collins (OCC model of emotions) [47] combined with the five factor model of personality [18] and a simulation of mood, to bias the emotions' intensities. All five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) influence the intensities of the different emotion types. We therefore adopted essential psychology research results on how personality influences emotions to achieve a more human-equivalent emotion simulation. Watson and Clark [64] and Becker [5] have empirically shown that personality, described

through the big-five traits, impacts the intensity of emotions. They discovered, e.g., that extravert people experience positive emotions more intensely than negative emotions. In our computational model this is realized by the change of an emotion's basic intensity, the so-called emotion intensity bias. Note that, the intensity of elicited emotions cannot be lower than the emotion intensity bias. When the personality is defined by a graphical user interface one can directly observe the impact on the emotions intensity bias, see Fig. 12.

Figure 12 consists of two screen shots showing the direct impact of the change of the extravert personality trait on emotions' intensity bias. In the example the extravert trait value is increased by moving the slider to the right side. As a consequence the basic emotion intensities of positive emotions increase. Note that not all emotions are biased in the same way. This depends on the fact that personality traits potentially bias emotion intensities at different strengths. Also the intensity biases are influenced by a person's current mood, see next section. The OCC cognitive model of emotions is based on the concepts of appraisal and intensity. The individual is said to make a cognitive appraisal of the current state of the world. Emotions are defined as valenced reactions to events of concern to an individual, actions of those

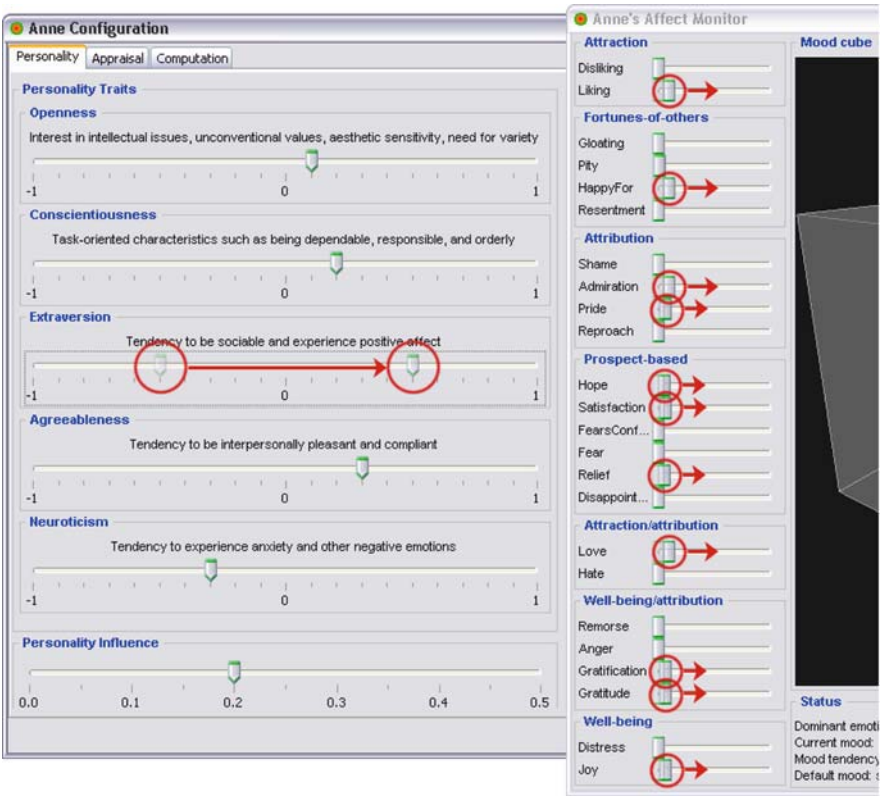


Fig. 12 Impact of personality traits on emotion intensities

she/he considers responsible for such actions and objects/persons. ALMA is able to compute all 24 emotions that are defined by the OCC theory.

The employed computational model of moods is based on the psychological model of mood (or temperament) proposed by Mehrabian [38]. Mehrabian describes mood with the three traits pleasure (P), arousal (A), and dominance (D). Each trait represents a specific mood component. Pleasure describes how much an individual enjoys the actual situation. Arousal stands for the excitement of an individual in the actual situation. Dominance describes to what extent an individual controls the actual situation. The three traits are nearly independent, and form a three-dimensional mood space. A PAD mood can be located in one of eight mood octants. A mood octant stands for a discrete description for a mood: +P+A+D is exuberant, −P−A−D is bored, +P+A−D is dependent, −P−A+D is disdainful, +P−A+D is relaxed, −P+A−D is anxious, +P−A−D is docile, and −P+A+D is hostile. Generally, a mood is represented by a point in the PDA space.

For a mood computation, it is essential to define a person's default mood. A mapping, empirically derived by Mehrabian [20], defines a relationship between the big five personality traits and the PAD space. The big-five traits can be obtained by a UbisWorld user profile. If they are not defined, a neutral mood will be assumed.

We define the mood strength by its distance to the PAD zero point. The maximum distance is $\sqrt{3}$. This is divided into three equidistant sections that describe three discrete mood intensities: slightly, moderate, and fully. Using the above mentioned mapping and the mood strength definition, a person whose personality is defined by the following big five personality traits: openness = 0.4, conscientiousness = 0.8, extraversion = 0.6, agreeableness = 0.3, and neuroticism = 0.4 has the default mood slightly relaxed (pleasure = 0.38, arousal =–0.08, dominance = 0.50).

An AffectMonitor, shown in Fig. 13, is used to visualize a person's current mood and mood changing emotions. The left side of the AffectMonitor shows all emotions and their intensities. Newly elicited emotions are marked dark gray (red). The right side shows a three-dimensional PDA mood cube displaying the current mood (the highlighted octant stands for the discrete mood description, whereas the light gray (yellow) ball reflects the actual mood) and all active emotions (dark gray (red) balls). Below, the affective state, including the current dominant emotion, and the default as well as the current mood, is displayed. The current mood also influences the intensity of active emotions. The theory is that the current mood is related to personality values that interfere with a person's personality values. Based on the current mood, the most intense related personality trait is identified. The actual value of this trait blends over the person's original personality trait value and is used to regulate the intensity of emotions. This increases, for example, the intensity bias of joy and decreases the intensity bias of distress, when a person is in an exuberant mood.

9.2.1 Mood Changes

According to Morris [39, p. 24] conditions for mood changes can be divided into (a) the onset of a mildly positive or negative event, (b) the offset of an emotion-inducing event, (c) the recollection or imagining of an emotional experience, and (d) the

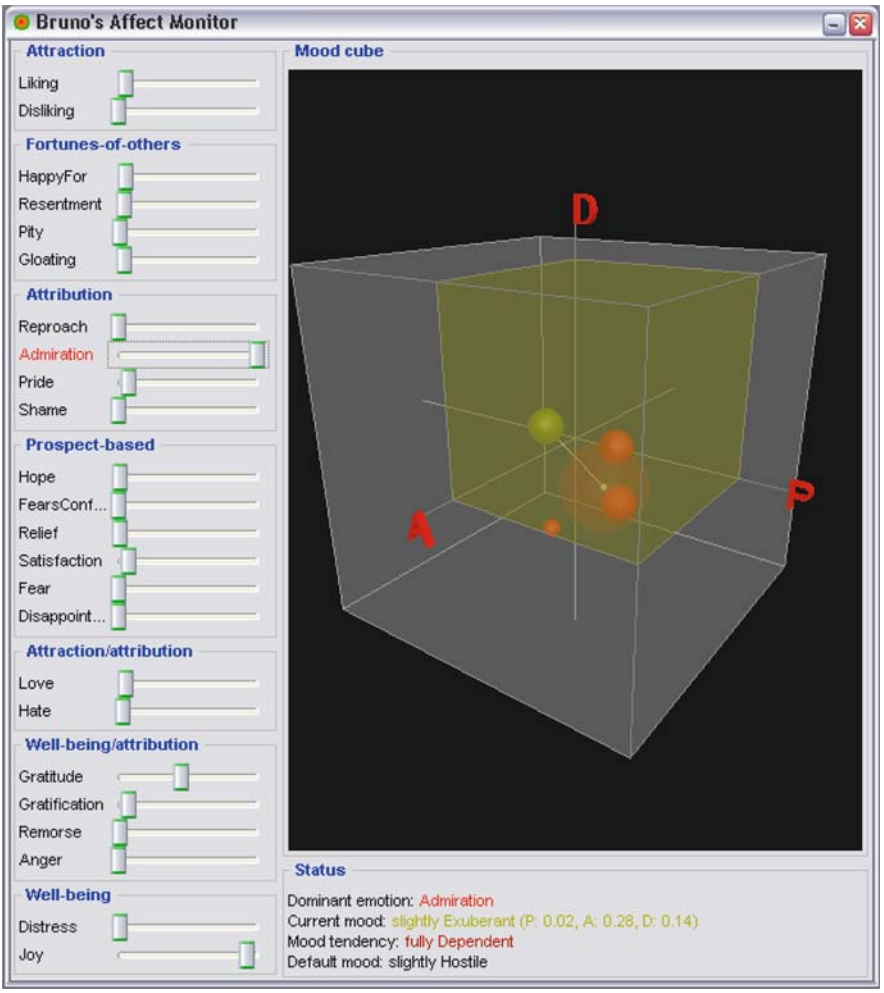


Fig. 13 AffectMonitor visualizes a person's mood and emotions

inhibition of an emotional response in the presence of an emotion-inducing event. To keep the modeling of mood changes as lean as possible, we take elicited emotions as the mood changing factor. In order to realize this, emotions must be somehow related to a specific mood. While using the PAD space for modeling mood, it is obvious to relate emotions to the PDA space too.

We rely on Mehrabian's mapping of emotions into the PAD space [38]. However, not all 24 emotion types of the OCC-Modell are covered by this mapping. For those that lack a mapping, we provide the missing pleasure, arousal, and dominance values by exploiting similarities to comparable emotion types [25].

Our approach to the human-like simulation of mood changes relies on a functional approach. We concentrate on how the intensity of emotions influences the change of the current mood. Moreover, we consider the aspect that the more expe-

riences a person makes that support a specific mood, the more intense this person's mood gets. For example, if a person's mood can be described as slightly anxious and several events let the person experience the emotion fear, the person's mood might change to moderate or fully anxious.

9.2.2 Appraisal Based Affect Computation

In our cognitively inspired affect computation, the first step is to evaluate relevant input by imitating a person's own subjective appraisal of a current situation. The situation is appraised according to two different concepts: (1) situational appraisal: events, actions, and objects and (2) interaction appraisal. The appraisal is realized by specific tags that relates to the appraisal concepts: (1) basic appraisal tags and (2) act appraisal tags. All appraisal tags are defined in a person's ubis world ontology.

Basic appraisal tags express how a person appraises the event, action, or object in the current focus. There are 12 basic tags for appraising events, e.g., the tag GoodEvent, which marks an event to be good according to the subjective view of the one which does the appraisal. The other event tags are BadEvent, GoodEventForBadOther, GoodEventForGoodOther, BadEventForGoodOther, BadEventForBadOther, GoodLikelyFutureEvent, GoodUnlikelyFutureEvent, BadLikelyFutureEvent, BadUnlikelyFutureEvent, EventConfirmed, and EventDisconfirmed. For appraising actions, there are four basic appraisal tags: GoodActSelf, BadActSelf, GoodActOther, and BadActOther. And finally there are two basic tags for appraising objects: NiceThing and NastyThing. All basic appraisal tags together are the basic set of a high-level appraisal language which can be used for a subjective appraisal of situations. These tags can be used to appraise dialog acts and other affective signals. For each of these types the appraisal language provides specific tags: act appraisal tags and affect display appraisal tags. Act appraisal tags represent the underlying communicative intent of an utterance, e.g., tease or congratulate.

Generally, the output of the appraisal process is a set of emotion eliciting conditions. Based on them active emotions are generated that in turn influence a person's mood. On the technical side, each person has their own ALMA process, which processes affect input. The input consists of appraisal tags, dialog act input, emotion and mood input, information about who is speaker, addressee, and listener. The computed affect (emotions and mood) is then passed to the other modules of the application.

The evaluation of this computational model of affect shows that nearly all affect types are plausibly represented in dialog scenarios.

10 Conclusions

The project BAIR has been concerned with research about user adaptation in instrumented rooms, with user-centered approaches in respect of the limitation of cognitive and technical resources. One focus was set on the role of cognitive and affective states of the user for generating affective responses from the instrumented

environment and for adapting the presentation of information. We developed an affective layer between the user services and the physical environment. In BAIR, we investigated the introduced concepts and novel interaction methodologies for proactive and user-centered support in multi-user instrumented environments. Research findings were also used in collaboration with other projects.

Acknowledgments The research project BAIR was supported by the German Research Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Project EM 4. We also would like to thank the Transfer Unit on Cognitive Technologies for Real-Life Applications, TFB53, Project TB2, RENA for their cooperation in the research on indoor navigation solutions.

References

1. Aaker, J. Dimensions of brand personality. Journal of Marketing Research, 34(3):342–352 (1997).
2. André, E., Klesen, M., Gebhard, P., Steve Allen, T.R. Integrating models of personality and emotions into lifelike characters. In A. Paiva, C. Martinho (Eds.), Proceedings of the Workshop on Affect in Interactions – Towards a New Generation of Interfaces in Conjunction with the 3rd i3 Annual Conference (pp. 136–149). Italy: Siena (1999).
3. Arons, B. A review of the cocktail party effect. Journal of the American Voice I/O Society, 12:35–50 (1992).
4. Barrington, L., Lyons, M., Diegmann, D., Abe, S. Ambient display using musical effects. In IUI '06: Proceedings of the 11th International Conference on Intelligent User Interfaces (pp. 372–374). ACM Press, New York, USA (2006).
5. Becker, P. Structural and relational analyses of emotion and personality traits. Zeitschrift für Differentielle und Diagnostische Psychologie, 22(3):155–172 (2001).
6. Blattner, M., Sumikawa, D., Greenberg, R. Earcons and icons: Their structure and common design principles. Human Computer Interaction, 4:11–44 (1989).
7. Brandherm, B. Eingebettete dynamische Bayessche Netze n-ter Ordnung. PhD thesis, Computer Science Institute, Saarland University, Germany (2006).
8. Brandherm, B., Schwartz, T. Geo referenced dynamic bayesian networks for user positioning on mobile systems. In Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, volume 3479/2005 of Lecture Notes in Computer Science (pp. 223–234). Springer-Verlag Berlin Heidelberg, Munich, Germany (2005).
9. Brandherm, B., Schwartz, T. Geo referenced dynamic Bayesian networks for user positioning on mobile systems. In T. Strang, C. Linnhoff-Popien (Eds.), Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, volume 3479/2005 of Lecture Notes in Computer Science (pp. 223–234). Springer-Verlag Berlin Heidelberg, Munich, Germany (2005).
10. Bregman, A. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press (1990).
11. Brewster, S. Using non-speech sounds to provide navigation cues. ACM Transactions on Computer–Human Interaction, 5:224–259 (1998).
12. Bruner, G.C. Music, mood, and marketing. Journal of Marketing, 54:94–104 (1990).
13. Butz, A., Jung, R. Seamless user notification in ambient soundscapes. In IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces (pp. 320–322). ACM Press, New York, NY, USA (2005).

14. Buxton, W., Gaver, W., Bly, S. Tutorial chapter 6: The use of non-speech audio at the interface. In Proceedings of Computer Human Interaction (CHI). ACM Press; Addison-Wesley, New Orleans (1991).
15. Camurri, A., Leman, M. Gestalt-based composition and performance in multimodal environments. In Joint International Conference on Cognitive and Systematic Musicology (pp. 495–508) (1996).
16. Cherry, E.C. Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustic Society of America, 25:975–979 (1953).
17. Costa, P., McCrae, R. The Neo Personality Inventory Manual. Odersa, FL: Psychological Assessment Resources (1985).
18. Davidson, R. On Emotion, mood, and related affective constructs. The nature of Emotion: Fundamental Questions (pp. 51–55). New York: Oxford University Press (1994).
19. Davies, J. The Psychology of Music. London: Hutchinson (1978).
20. Duggan, B., Deegan, M. Considerations in the usage of text to speech (tts) in the creation of natural sounding voice enabled web systems. In ISICT '03: Proceedings of the 1st International Symposium on Information and Communication Technologies (pp. 433–438). Trinity College Dublin, Ireland (2003).
21. Eysenck, M., Keane, M. Cognitive Psychology: A Student's Handbook. Hove: Psychology Press (2005).
22. Feld, M. Erzeugung von Sprecherklassifikationsmodulen für multiple Plattformen, Diploma Thesis, Soarland University (2006).
23. Fiske, S., Taylor, S. Social Cognition. New York: McGraw-Hill (1991).
24. Garofolo, J. e. A. DARPA TIMIT CD-ROM: An Acoustic Phonetic Continous Speech Database. Gaithersburg, MD, USA: National Institute of Standards and Technology (1998).
25. Gebhard, P. Alma – a layered model of affect. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh, M. Wooldridge (Eds.), Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (pp. 29–36) (June 2005).
26. Gill, A.J., Oberlander, J., Austin, E. The perception of e-mail personality at zero-acquaintance. Personality and Individual Differences, 40:497–507 (2006).
27. Heckmann, D. Ubiquitous user modeling. PhD thesis, Department of Computer Science, Saarland University, Germany, DISKI 297, ISBN 3898382974, Aka Verlag (2005).
28. Heckmann, D., Schwartz, T., Brandherm, B., Kröner, A. Decentralized User Modeling with UserML and GUMO (pp. 61–66). Scotland: Edinburgh (2005).
29. Johanson, B., Fox, A. The event heap: A coordination infrastructure for interactive workspaces. In Proceedings of the Workshop on Mobile Computing Systems and Applications (2002).
30. Jung, R., Butz, A. Effectiveness of user notification in ambient soundscapes. In Proceedings of the workshop on Auditory Displays for Mobile Context-Aware Systems at Pervasive 2005 (pp. 47–56). Munich, Germany (2005).
31. Jung, R., Heckmann, D. Ambient audio notification with personalized music. In Workshop on Ubiquitous User Modeling at ECAI 2006 (pp. 16–18). Riva del Garda, Italy (2006).
32. Jung, R., Schwartz, T. A location-adaptive human-centered audio email notification service for multi-user environments. In J.A. Jacko (Ed.), Human–Computer Interaction, vol. 4552 of LNCS (pp. 340–348). New York: Springer (2007).
33. Jung, R., Schwartz, T. Peripheral notification with customized embedded audio cues. In Proceedings of the 13th International Conference on Auditory Displays (pp. 221–228). Schulich School of Music, McGill University, Montreal, Canada (2007).
34. Krause, R. Affekt, Emotion, Gefühl (2nd edn., pp. 30–36). Stuttgart: Kohlhammer (2002).
35. Legaspi, R., Hashimoto, Y., Moriyama, K., Kurihara, S., Numao, M. Music compositional intelligence with an affective flavor. In Proceedings of Conference on Intelligent User Interfaces (IUI) (2007).
36. McCrae, R., John, O. An introduction to the five-factor model and its applications. Journal of Personality, 60:175–215 (1992).

37. McRae, R., John, O. An introduction to the five-factor model and its applications. Journal of Personality, 60:175–215 (1992).
38. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology: Developmental, Learning, Personality, Social, 14:261–292 (1996).
39. Morris, W.N. Mood The Frame of Mind. New York: Springer (1989).
40. Müller, C. Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]. PhD thesis, Computer Science Institute, University of the Saarland, Germany (2005).
41. Müller, C., Feld, M. Towards a multilingual approach on speaker classification. In Proceedings of the 11th International Conference "Speech and Computer" SPECOM 2006 (pp. 120–124). Anatolya Publishers, St. Petersburg, Russia (2006).
42. Nass, C., Isbister, K., Lee, E.-J. Truth is beauty: Researching embodied conversational agents. Embodied conversational agents (pp. 374–402). Cambridge, MA: MIT Press (2000).
43. Nijholt, A., Rist, T., Tuijnenbreijer, K. Lost in Ambient Intelligence? Panel Session. In Proceedings of CHI'04 (pp. 1725–1726). ACM, New York (2004).
44. North, A., MacKenzie, L., Hargreaves, D. The effects of musical and voice "fit" on responses to advertisements. Journal of Applied Social Psychology, 34(8):1675–1708 (2004).
45. Nowson, S. The langauge of weblogs: A study of genre and individual differences. PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics. (2006).
46. O'Conaill, B., Frohlich, D. Timespace in the workplace: dealing with interruptions. In CHI '95: Conference Companion on Human Factors in Computing Systems (pp. 262–263). ACM Press, New York, NY, USA (1995).
47. Ortony, A., Clore, G.L., Collins, A. The Cognitive Structure of Emotions. Cambridge, MA: Cambridge University Press (1988).
48. Pennebaker, J., King, L. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77:1296–1312 (1999).
49. Pinhanez, C. The everywhere displays projector: A device to create ubiquitous graphical interfaces. Lecture Notes in Computer Science (2001).
50. Reeves, B., Nass, C. The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge, MA: CSLI Publications and Cambridge university press (1996).
51. Reybrouck, M. Gestalt concepts and music: Limitations and possibilities. In Joint International Conference on Cognitive and Systematic Musicology (pp. 57–69) (1996).
52. Scherer, K., Zentner, M. Music and Emotion: Theory and Research (Chapter 16, pp. 361–392). Oxford, England: Oxford University Press (2001).
53. Schiel, F. Speech and speech-related resources at BAS. In Proceedings of the First International Conference on Language Resources and Evaluation (pp. 343–349). Granada, Spain (1998).
54. Schmitz, M., Butz, A. Safir: Low-cost spatial audio for instrumented environments. In Proceedings of the 2nd International Conference on Intelligent Environments. Athens, Greece, (2006).
55. Schmitz, M., Krüger, A., Schmidt, S. Modelling personality in voices of talking products through prosodic parameters. In Proceedings of the 10th International Conference on Intelligent User Interfaces (pp. 313–316) (2007).
56. Schneider, M. A smart shopping assistant utilising adaptive plan recognition. In Proceedings of the Workshop 'Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen' (ABIS) at LLWA (2003).
57. Schröder, M., Grice, M. Expressing vocal effort in concatenative synthesis. In Proceedings of the 15th International Conference of Phonetic Sciences (pp. 2589–2592) (2003).
58. Schwartz, T., Brandherm, B., Heckmann, D. Calculation of the user-direction in an always best positioned mobile localization system. In Proceedings of the International Workshop on Artificial Intelligence in Mobile Systems (AIMS). Salzburg, Austria (September 2005).

59. Spassova, L. Fluid Beam – A Steerable Projector and Camera Unit. Student and Newbie Colloquium at ISWC/ISMAR (2004).
60. W3C-EMMA. EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft, 9 April 2007, http://www.w3.org/TR/emma/, last accessed: 31.10.2007 (2007).
61. Wasinger, R. (Ed.) Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations. Akademische Verlagsgesellschaft, Berlin, Germany (2006).
62. Wasinger, R., Krüger, A., Jacobs, O. Integrating intra and extra gestures into a mobile and multimodal shopping assistant. In Proceedings of the 3rd International Conference on Pervasive Computing (pp. 297–314) (2005).
63. Wasinger, R., Krüger, A., Jacobs, O. Integrating intra and extra gestures into a mobile and multimodal shopping assistant. International Conference on Pervasive Computing (2005).
64. Watson, D., Clark, L.A. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. Journal of Personality, 2(60): 441–476 (1992).

Seamless Resource-Adaptive Navigation

Tim Schwartz, Christoph Stahl, Jörg Baus, and Wolfgang Wahlster

1 Introduction

Research in the project RENA (REsource-Adapative NAvigation) together with DFKI GmbH, BMW Research and Technology AG, and Eyeled GmbH has been concerned with the conceptual and methodological foundations and the design of a resource-adaptive platform for seamless outdoor and indoor navigation that can serve as a basis for product development by the companies in the RENA consortium. Future in-car assistance systems will have a user interface, which adapts to the driver's current exposure caused by the actual traffic situation. Based on concepts developed in [15] such in-car assistance systems will use the car's serial-production sensory equipment to detect the driver's momentary cognitive load without the additional use of biosensors attached to the driver. In case the system detects that the driver has high cognitive load or is in stress, system messages, which are not time-critical, like directions from the navigation system or incoming telephone calls will be delayed until the driver's cognitive load decreases. Besides these new features, in-car assistance systems will use car-2-car and car-2-X communication capabilities to exchange information about the traffic situation and road conditions with other cars and their drivers. Such information gained from the car's sensory equipment, e.g., rain and moisture sensors, sensors from the ABS, and/or sensors from airbags will be mapped onto concepts defined in local danger warning ontologies to realize local danger warnings (see [5, 8]). In parking garages, where the car's assistance system cannot receive GPS signals, the car's position is determined using the car's build-in gyroscope as described in [18] or on the basis of active RFID positioning (see [12]).

The distinguishing feature of RENA's ubiquitous navigation service is its adaptability to the user, the situation, and the technical resources. Various positioning and wireless communication technologies are combined synergistically within the developed platform. Due to the fact that those aforementioned theses are still under nondisclosure agreement up to the time of preparing this chapter, we will focus

T. Schwartz (✉)
DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany
e-mail: Tim.Schwartz@dfki.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_11,

on YAMAMOTO (Yet Another MAp MOdeling TOol), a map modeling toolkit for indoor pedestrian navigation, and our positioning system LORIOT (Localization and ORientation for Indoor and Outdoor EnvironmenTs). Both systems serve as the core components for different applications such as route visualization and a location-based To-Do Organizer.

The remainder of the chapter is structured as follows: In the next section we will review two navigation systems, which serve as starting points for the later developments. Then, we will give a short overview on the overall system's architecture followed by detailed descriptions of the core components YAMAMOTO and LORIOT. The next section will introduce applications that use both components in order to help users solve different tasks in the navigation domain. The last section will summarize the results.

2 REAL and BPN as the Basis of our Extensions

In this section we provide a brief review of two navigation systems which were developed within our project and served as the basis for further developments, which are described in the remainder of the chapter.

The REAL system was a hybrid pedestrian navigation aid that could tailor its graphical route descriptions to limited technical resources of different devices and the user's limited cognitive resources. Based on the introduced concept of active and passive location sensitivity, a navigation aid for pedestrians was developed. Active location sensitivity delegates the computational burden to the mobile device. The mobile device actively detects the actual location on its own. According to this information the presentation of graphical route descriptions is generated automatically to comply with the user's requests. In contrast to this, passive location sensitivity reallocates nearly all necessary computations to the instrumented environment. The mobile device just passively presents the information that it receives from senders in its local environment, specifically information prepared for the actual location. The navigation aid integrated three subsystems: (i) a stationary information kiosk, (ii) IRREAL (Infrared REAL) for indoor navigation tasks based on infrared senders, and (iii) ARREAL (Augmented Reality REAL) for outdoor navigation based on GPS-satellites. IRREAL uses passive location sensitivity, whereas ARREAL relies on the active counterpart. The whole system has been designed in such a way that the changeover between both adaptation paradigms is barely noticeable for the user. For a more detailed look on the system please refer to [3, 4].

In cooperation with BMW AG, the BMW Personal Navigator (BPN) was developed. The system combines a desktop event and route planer, a car navigation system, and a multi-modal, in-, and outdoor pedestrian navigation system for a PDA. BPN offers a situated personalized navigation service and seamlessly integrates 2D and 3D maps with speech in- and output on the mobile device. With its research focus on multi-modal interaction, BPN allows the user to interact with the navigational map through the combined use of speech (English and German) and stylus gesture. We investigated three different types of situations in which navigational services may be of interest. We aimed at providing a service that transparently

combines the desktop PC at home, a built-in car navigation system, and a PDA. Such a situated personalized navigation service allows travelers to prepare their trips at home in order to obtain route directions and maps, choose personalized events of interest at their destination, book an appropriate hotel, and retrieve further information, like the current weather. This information is collected and stored in a travel itinerary for each trip. With the help of the PDA, travelers can also make use of their travel itinerary in the car and as pedestrians on-foot, both inside and outside of buildings. At home the desktop PC is used to make all travel arrangements provided by a personal navigation server that can be accessed over the Internet. The travel itinerary is then synchronized with the PDA which then allows access to the travel itinerary without the need for a direct Internet connection. In the car, the PDA connects locally to the car navigation system, which in turn allows to transfer the travel itinerary from the PDA to the car navigation system. During the navigation task in the car, the PDA remains predominantly silent, and the car navigation system takes control in guiding the traveler to the selected destination. Before leaving the car, the PDA receives the actual park spot coordinates, which are added to the travel itinerary and may help to find a way back to the car later. For the pedestrian, the PDA plays a much more vital role. It displays the 3D map information included in the travel itinerary and guides the traveler with verbal and graphical route directions. It can also be used to store geo-referenced user data (e.g., voice memos) and respond to multi-modal requests regarding landmarks in the environment as described in [26, 14].

3 Overall System Architecture of the New Navigation Framework

Based on the results of the aforementioned prototypes, we will now introduce the components of the new navigation framework and the way they interact with each other as illustrated in Fig. 1. The complete navigation framework consists of different services which run on workstations, servers, and mobile devices. Furthermore, our lab is instrumented with active RFID tags and infrared beacons for positioning purposes, and Bluetooth-enabled public displays for location and user adaptive information services. In more detail, the following base applications comprise the software environment:

YAMAMOTO is an application that was developed for easy and rapid modeling of indoor and outdoor environments. YAMAMOTO is also able to calculate routes from one room to another in one building or from a room in one building to a room in another building.

LORIOT is an always best positioned (ABP) system which uses active RFID tags and infrared beacons as well as GPS to calculate the position of users in indoor and outdoor environments. It runs on PDAs and has the highest precision of all positioning systems described in this chapter. A user can choose to reveal their position to the instrumented environment or protect their privacy.

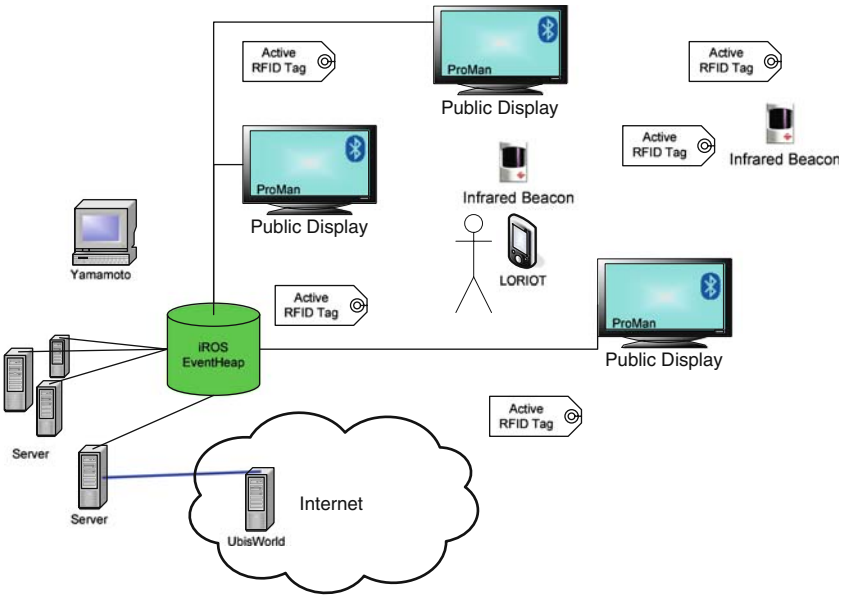


Fig. 1 Overview of all components discussed in this chapter, and how they are interconnected

UbisWorld is a ubiquitous user modeling service that provides user models as well as spatial ontologies. An important feature of UbisWorld is the so-called UbisNames which provide a unique symbolic identifier for each entity (people, places, objects) that is modeled by the service. A more detailed description of UbisWorld can be found in [24] in this volume.

iROS EventHeap The iROS (Interactive Room Software Infrastructure) EventHeap [13] is a part of the Stanford Interactive Workspaces Project. We use the EventHeap to interconnect the different services and applications that either run on servers, workstations, PDAs, or mobile phones. Each application or service can register at the EventHeap and can then post and receive messages (events).

In the following we will provide a detailed view on the core components YAMAMOTO in Sect. 4 and LORIOT in Sect. 5.

4 Providing Map Material for Pedestrian Navigation

Our research aims toward a system that provides ubiquitous navigational aid for its users, with emphasis on indoor environments, but which also covers outdoor places and routes on a large scale. As mentioned in Sect. 2 this vision of a seamless indoor–outdoor navigation system has already been implemented in the REAL and BPN research prototypes (see [3, 14, 4]). Whereas these prototypes have been quite

successful in demonstrating the overall concept, we have learned several lessons about location modeling. The BPN system is based on a commercial street database, which has been imported into an open source GIS system and manually extended by indoor route segments. This approach lacks the scalability that would be required for a truly ubiquitous location model, as too many buildings need to be represented and maintained in a single database.

Since the environment model is solely based on a path network which is represented as a graph of line segments in two dimensions, it has severe limitations for the pedestrian navigation domain. The situation of a pedestrian differs from driving tasks, since the user is not bound to follow paths or streets. Instead users typically cross open spaces directly following their line of sight. The model should particularly reflect this and represent places as polygonal regions.

Furthermore, since the GIS system's routing module operates with 2D geographic coordinates, it takes several workarounds to denote destinations in the upper floors of buildings. Inside the building, no GPS data can be received, so the BPN system used infrared beacons, and the position of the user was looked up from a database that returned the geometric coordinates of the received beacon IDs (16 bit of information). The installation of the beacon infrastructure took several days, since no tools were available at this time to graphically model the position, range, and orientation of the beacons. Other difficulties arose from the fact that three different location models were required. Start and destination addresses are usually given as geographic locations (postal addresses) and have to be mapped to physical locations in the WGS84 [7] coordinate system (longitude, latitude) used by GPS. These geographical coordinates have to be mapped to screen coordinates (x, y) in the map's texture bitmap reference system in order to visualize the position of the user. Indoors, no WGS84 coordinates are known, and the paths and beacons have been entirely modeled in bitmap coordinates instead. The alignment of the indoor space with the outdoor space has been done manually and hence was error prone. Besides map visualization, the BPN system has been designed to convey automatically generated verbose instructions to the user, such as "turn right after 10 m." As the underlying location model consisted of a set of two-dimensional floor maps without height information, additional annotation workarounds were required to guide the user through a staircase, e.g., "Please go up the stairs to the 2nd floor."

The commercial providers of navigational maps for mobile systems have recognized the benefit of 3D visualizations, but they are still focused on outdoor environments. As pedestrians spend most of their time inside buildings, indoor environments need to be modeled in 3D with multiple floors and landmarks. Indoor's decision points are more complex than outdoor's because stairs and elevators add choices. For the same reason, routes cannot be depicted easily in a single map, so that indoor wayfinding tasks generally pose a high cognitive load to the user.

In summary, we conclude the following research issues for pedestrian navigation in mixed indoor/outdoor environments from our previous experiences with our research prototypes:

- scalability and maintainability of the underlying location model
- polygonal representation of regions instead of abstract line segments
- mapping of beacons for indoor positioning
- hybrid geographic (symbolic) and physical (geometric) location modeling
- modeling height information.

During the course of our project, we have continuously developed our own map modeling toolkit, which is called YAMAMOTO. The toolkit is positioned between proprietary 2D location models that are typical for ubiquitous and pervasive computing research projects on indoor navigation and professional 3D CAD (computer-aided design) tools for architects. CAD tools require a high level of experience; the designer has to manually cut out windows and doors from solid walls and take care about window sills, choices of door handles, or steps of a staircase. Such a high level of detail is not required for route finding and presentation purposes. Our approach strives to minimize the modeling effort. By following the motto to keep everything as simple as possible, we have intentionally reduced the degrees of freedom by half of a dimension in order to allow for a simpler and easier to learn user interface.

Now what exactly does this mean, and what are the implications? Rooms of a building are represented only by their outline as flat polygon objects. Each polygon object is defined by an ordered sequence of vertices. Each vertex is represented through Cartesian coordinates as a triple of (x, y, z) values. The z-value allows representing the room's height above ground level, so that multiple floors can be represented. Polygons can have several symbolic attributes, such as name, type, and accessibility for pedestrians. Polygons that are defined by vertices from two different levels represent connections such as ramps, stairs, or escalators. Figure 2 (left) shows an example, where the polygon "Corridor.14021" is defined as sequence of vertices with index (1, 2, 3, 4, 5, 6, 7, 8). In order to allow for route finding, it is important to know the semantics of connections between polygons. Thus each edge is attributed by their passability: edges that represent walls or windows are set to be "not passable"; in our example, edge (8, 1) represents a wall and edge (6, 7) connects the corridor with the adjacent staircase and is annotated to be "passable for pedestrians." On the right-hand side in Fig. 2, a sample path is shown that has been calculated based on start and end points within the 2.5D location model.

Based on the outlines of the rooms and some additional annotation of type and height, YAMAMOTO automatically creates the building structure in full 3D. By using the predefined building blocks shown in Fig. 3, edges can be visualized in the perspective views as walls, doors, murals, or handrails.

In addition to polygon objects, the mapping of navigational fixpoints, such as beacons or landmarks, is required for the indoor positioning system. For this reason we allow additional geometric primitives, such as points, spheres, and sections, which can be used to represent the position of the beacon or more precisely the reception range of the signal emitted by the beacon.

In YAMAMOTO one can choose among different viewpoints at any time. The orthogonal view shows a top-down projection of the model similar to traditional maps. The perspective view shows the model from an allocentric viewpoint outside

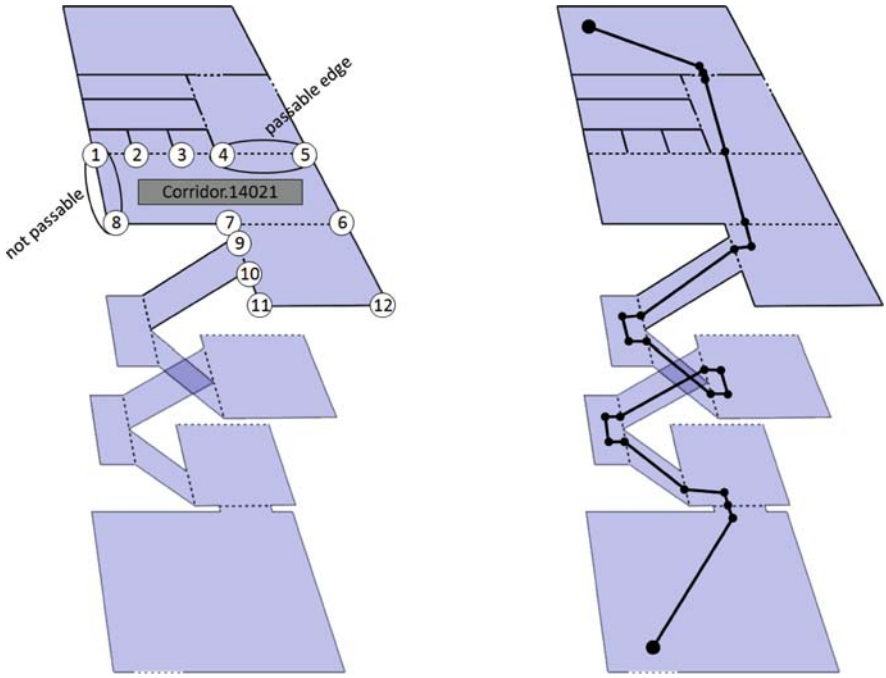


Fig. 2 The 2.5D data model (*left*) and a route between two points (*right*)

the model, as shown in Fig. 5, and allows for free rotation and zoom. The user itself can be virtually represented in the model through an avatar object. The egocentric perspective shows the model from the viewpoint of this avatar, see Fig. 5. It allows for the virtual exploration of the modeled environment. It also creates a demand for interior items that could serve as landmark objects for route descriptions. Rooms can be equipped with predefined 3D objects, like shelves, tables, or pictures, as depicted in Figs. 4 and 5.

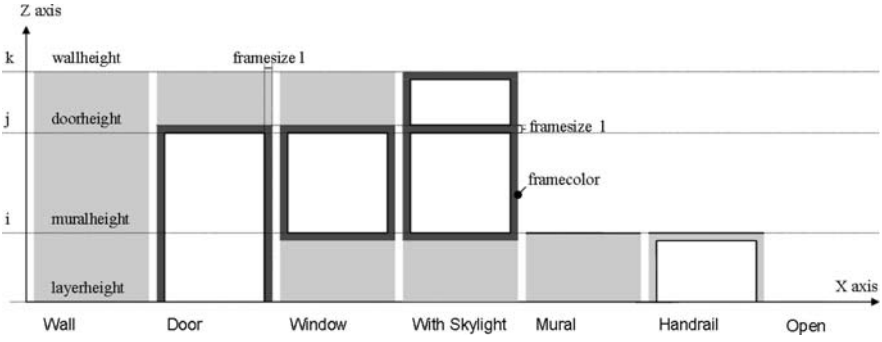


Fig. 3 Set of building blocks used to automatically create 3D geometry from the 2.5D model

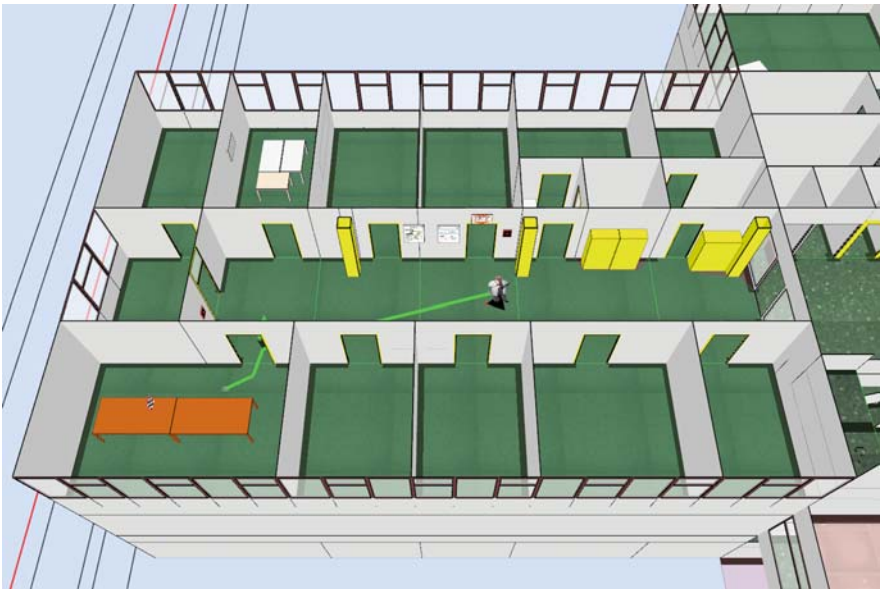


Fig. 4 Allocentric perspective

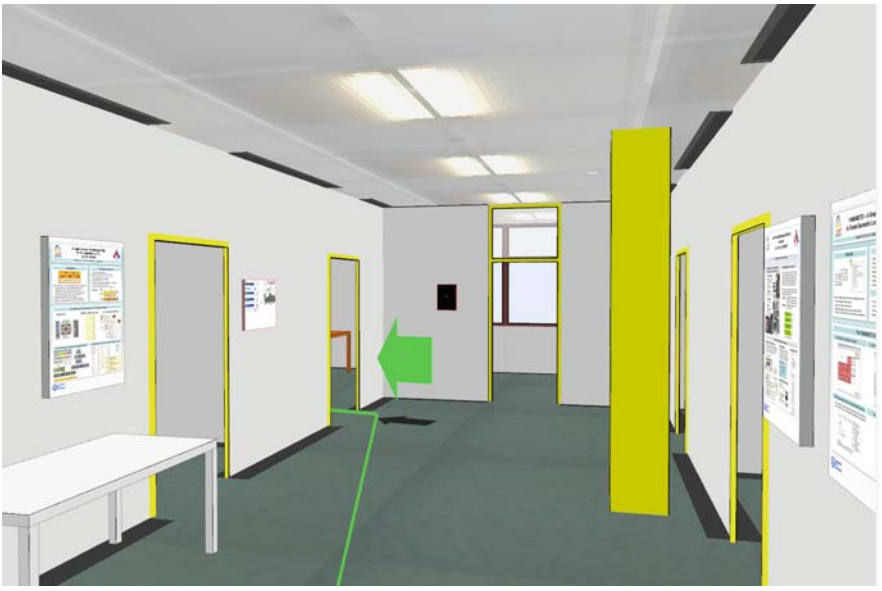


Fig. 5 Egocentric perspective

Furthermore, the toolkit allows to test the instrumentation of the environment with pervasive computing artifacts, i.e., beacons used for indoor positioning/navigation and public displays. The avatar view lets the designer virtually examine the visibility of the displays from various positions and helps to identify the best configuration, as described in [22]. In particular the possible interpretation of graphical signage, e.g., an arrow pointing upward, can be ambiguous depending on the actual context of the building. Such situations can be virtually evaluated and resolved before the signs are deployed.

As the toolkit has been designed with pedestrian navigation in mind, it includes a route finding module. It is able to generate routes between any two points in a model, which follow the line-of-sight whenever possible instead of a restrictive path network. Since even multi-story buildings can be represented as a single mesh, the pedestrians will be routed through staircases, if needed. During the modeling process, the results of the route finding module can be tested at any time within the editor.

5 The Always Best Positioned Paradigm

Today's navigation systems are designed to work for a specific platform in a well-defined environment, but they usually fail to work when the situation of the user is dynamically changing, e.g., when the user has to combine different means of transportation to reach a destination. One essential issue in this context is the switch between different positioning technologies. A navigation system intended to support its mobile users has to seamlessly cope with this technical problem and must be able to adapt to different technical constraints.

5.1 Exocentric and Egocentric Localization

Robust and resource-adaptive positioning is critical for the success of pedestrian navigation services. If the user is staying outside, localization can be done with the use of GPS. Unfortunately, GPS is not working in indoor environments, e.g., a shopping mall or an airport. There have been numerous attempts to overcome this restriction (e.g., [2, 25, 9]). All of these systems use some sort of senders (ultrasound beacons, infrared beacons, WiFi-hotspots, RFID tags, Bluetooth beacons, to name but a few) and corresponding sensors to detect or read these senders. Basically there are two options to set up such a system: installing sensors in the building and letting the user wear the sender or installing the senders in the environment and letting the user wear the sensors.

In the former case, the so-called exocentric localization (see left side of Fig. 6), the user is sending information to the environment and some centralized server uses these data to calculate their position. In other words, the user is tracked. In the latter

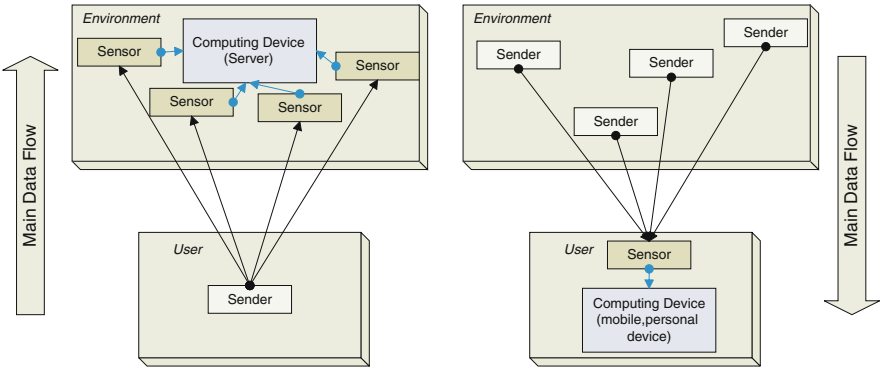


Fig. 6 Exocentric localization (*left*): the dataflow is from the user to the environment. Egocentric localization (*right*): the dataflow is from the environment to the user

case, the egocentric localization (see right side of Fig. 6), the user receives information from the environment and their personal device uses the data to calculate the current position.

5.2 LORIOT

In [6] and [21] we describe the basics of an egocentric localization system that uses Geo Referenced Dynamic Bayesian Networks (geoDBNs) to determine the position and orientation of a user.

Based on this technique we developed a seamless indoor/outdoor positioning system named LORIOT (Location and ORientation in Indoor and Outdoor environmenTs), which uses infrared beacons (IR beacons) and active RFID tags to determine the user's position in indoor environments and a GPS receiver for outdoor localization. LORIOT runs on Windows Mobile PDAs and uses the built-in infrared receiver, an attached active RFID reader, and a Bluetooth GPS receiver as sensors.

For indoor positioning the environment has to be instrumented with IR beacons and/or active RFID tags. These beacons and tags can be installed at the ceiling of the building and act as electronic landmarks. The reason to use two kinds of senders is that both technologies differ in their features, precision, and cost: IR beacons send out a 16 bit wide identification code. Due to the physical attributes of light, receiving such a beacon gives a very high probability that the user is standing near that particular beacon. Furthermore, if we know the direction of the infrared light beam, we can determine the user's direction. However, the disadvantage of IR beacons is that an IR sensor must be in the line of sight of the beacon and can thus be very easily blocked.

On the other hand, RFID tags send their information via radio waves, which can be received even when the PDA resides in the user's pocket. Due to reflections and damping of radio waves, receiving an RFID tag gives only little evidence that

the user is standing in its vicinity and the signal does not contain any directional information. Besides a unique, hardwired identification code, each active RFID tag contains a 56 byte wide, free accessible memory which we use to store its own geo-coordinates. These geo-coordinates, in the format of latitude, longitude, and elevation based on WGS 84 [7], are obtained with the help of our map modeling tool YAMAMOTO (see Sect. 4).

By combining different sensor readings with the help of dynamic Bayesian Networks, we achieve a positioning system that follows the above mentioned Always Best Positioned paradigm: As long as there are either IR beacons, RFID tags, or GPS satellites detectable, we will be able to estimate a position whose precision depends on the type of the sender. If we can receive all, we will get an even higher precision. The system adapts to the available resources (senders as IR beacons, RFID tags, and GPS satellites, or attached sensors as an infrared port, an active RFID reader card, or a GPS receiver) that can be used to find out about its own position.

In the following, we describe the idea of Geo Referenced Dynamic Bayesian networks and how they are used to accomplish a sensor fusion and cancel out false readings.

Bayesian networks (BNs) are a computational framework for the representation and the inference of uncertain knowledge. A BN can be represented graphically as a directed acyclic graph (see Fig. 7). The nodes of the graph represent probability variables. The edges joining the nodes represent the dependencies among them. For each node, a conditional probability table (CPT) quantifies these dependencies.

Dynamic Bayesian networks (DBNs) are an extension of Bayesian networks. With a DBN, it is possible to model dynamic processes: Each time the DBN receives new evidence a new time slice is added to the existing DBN. In principle, DBNs can be evaluated with the same inference procedures as normal BNs; but their dynamic nature places heavy demands on computation time and memory. To cut down these computational costs, it is necessary to apply so-called roll-up procedures that cut off old time slices without eliminating their influence on the newer time slices.

5.2.1 Estimation of the User Position

For the purpose of positioning, we let such a DBN represent the characteristics of the used senders. Figure 7 shows two geoDBNs: The one on the left was designed for indoor positioning only and thus contains only nodes for the IR sensor and the

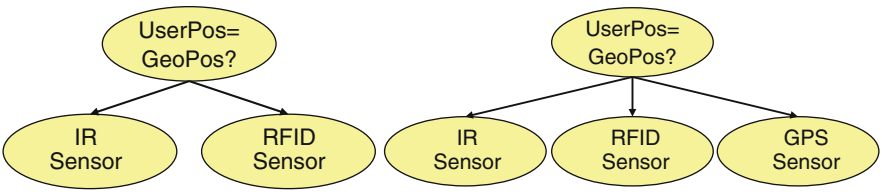


Fig. 7 *Right*: a simple geoDBN with IR- and RFID-Sensor nodes. *Left*: an extended geoDBN with additional GPS sensor node

RFID sensor. The geoDBN on the right is an extension, which also includes a node for a GPS sensor (GPS receiver) to enable seamless indoor/outdoor positioning. It is important to note that we do not use the DBN to represent all the senders that are physically installed in the environment. We use one small DBN that prototypically describes the reliability of the sender types (assuming that all senders of a certain type have the same reliability). This prototypical DBN is instantiated several times during the runtime of the system and each instantiation gets assigned to a geo coordinate *GeoPos*.

Figure 8 shows the network (with time slices) that we use in LORIOT. The top node at time slice t (labeled UserPos=GeoPos?) represents the probability that the user is standing at the assigned geo-coordinate *GeoPos*. The node to the left of it (UserPos=GeoPos?_1) represents the probability that was calculated in the previous time slice $t-1$. The bottom nodes (IRSensor, RFIDSensor, and GPSSensor) represent the probability that an IR beacon and/or an RFID tag installed at *GeoPos* can be detected under the condition that the user is standing at *GeoPos* and/or that the GPS receiver reports the coordinates of *GeoPos*, respectively.

Receiving an infrared signal gives very high evidence that the user is standing near the respective beacon (the infrared sensory data is nearly noise-free). Receiving an RFID tag gives smaller evidence that the user is standing near the tag, since the reader has a reading range of about 10 m and due to reflections of the radio waves, some RFID tags that are far away from the user can be detected. Modern GPS receivers are able to receive GPS signals within buildings if they are near to windows or outer walls, but the accuracy of the received GPS position is very low. Figure 9 shows the received GPS coordinates of a GPS receiver resting on a windowsill over a period of 25 min. In this set of data, the maximum deviation from the actual position is 444 m. Because of this high variation, the probability that the reported GPS position really coincides with the user's position is also very low.

The different characteristics of each technology are coded in the conditional probability tables (CPTs) of the IRSensor-, RFIDSensor-, and GPSSensor-nodes. The networks with their assigned coordinates are the geo-referenced dynamic Bayesian networks (geoDBNs). Each geoDBN represents the believe that the user is standing at the associated coordinate.

As stated above, the active RFID tags have a small internal memory that can be used to read and write data. We use this memory to store the coordinate of

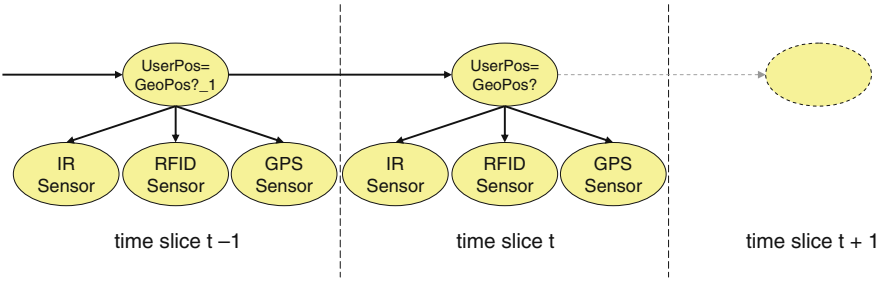


Fig. 8 A geoDBN with different time slices, which represent the different measurements at sequential points in time

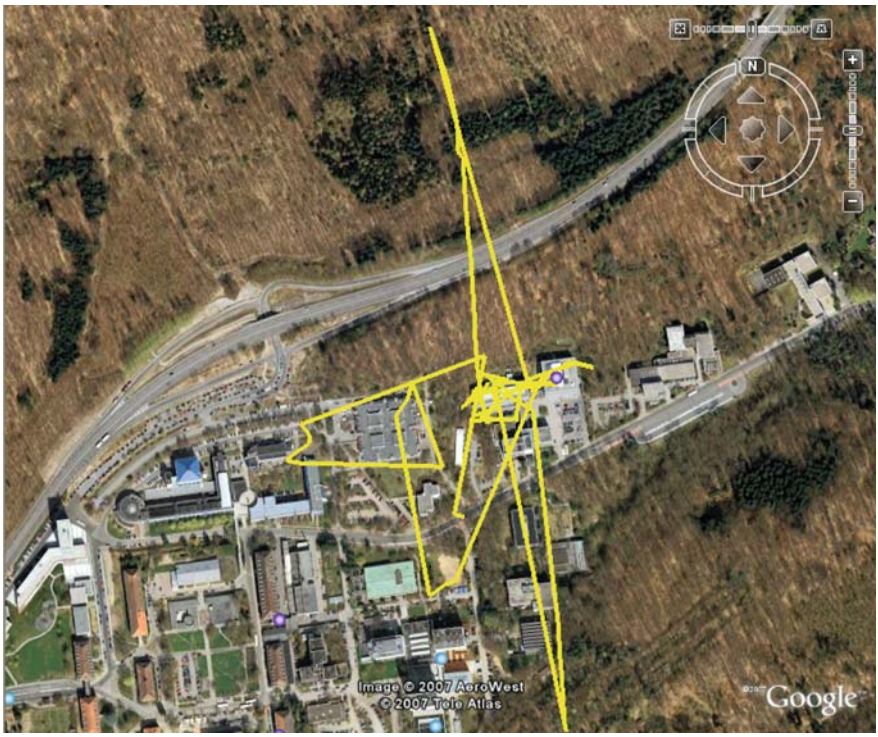


Fig. 9 Deviation of a GPS receiver resting on a windowsill inside a building

the tag. Since IR beacons only send a 16 bit wide identification code, the system needs a database that contains the associations between identification code and the respective geo-coordinate of the beacon. We also store these associations into the RFID tags, together with the information of the lighting direction of the IR beacon, so that the database can be distributed over several RFID tags in the environment. That way, the system can retrieve all the information it needs to calculate its own position from the environment. When a tag, a beacon, or a GPS position is sensed by the PDA, geoDBNs are instantiated and associated with the appropriate coordinates.

The calculation of the user position is done with a weighted sum of the received (or sensed) coordinates, where the probabilities of each geoDBN are used as weights:

$$\text{UserPos}(t) = \sum_{i=1}^{n} \alpha \; w(\text{GeoDBN}[i])\; \text{GeoPos}(\text{GeoDBN}[i]). \tag{1}$$

Here n is the number of existing geoDBNs at time t ($n \geq \#ReceivedSenders_t$), GeoPos(GeoDBN[$i$]) is the coordinate and w(GeoDBN[i]) the weight of the ith geoDBN. α is a normalization factor that ensures that the sum of all weights multiplied with α is 1.

To reduce calculation cost and memory usage the number of instantiated geoDBNs must be as low as possible. To achieve this goal, geoDBNs with a weight lower than $threshold_{use}$ are marked as unused (these geoDBNs provide only little evidence that the user is in the vicinity of their geo-coordinate). This threshold should match the a priori probability for the geoDBN at its first instantiation. To cope with resource restrictions, a maximum number of possible geoDBNs can be specified. If this number is exceeded, those geoDBNs that provide the lowest estimation will be deleted. To keep the overhead for memory management low or to prevent garbage collection – if the system is implemented in languages like Java or C# – geoDBNs that are marked as unused can be "recycled" by resetting them to initial values and new coordinates.

5.2.2 Orientation Estimation

Besides the information about the position of a user, the information about their walking or seeing direction can also be valuable. A common example is an electronic museum guide that not only needs to know where the visitor is standing but also which exhibit she is currently looking at, so that the system can give the respective explanations. The following describes how the direction information from both sensor types, infrared and active RFID, can be fused together again with the help of a dynamic Bayesian Network.

5.2.3 Orientation Information Through Infrared Beacons

The emitted infrared beam of our IR beacons has a range of about 6 m and a conical transmission characteristic due to the physical attributes of light.

Because of this conical transmission characteristic, the infrared beam is highly directional and the calculation of the walking direction is fairly easy. If the beacon sends its light in direction vector **v** (see Fig. 10) and the user receives the beacon then they are walking in direction $\mathbf{dir_{IR}} = -\mathbf{v}$. Of course this is only an estimation since the user can be slightly to the left or right of the main direction **v** due to the conical transmission characteristic. Also note that it is sufficient to use a 2D vector (the projection of the 3D direction vector – that includes the tilt of the beacon – on the x–z-plane).

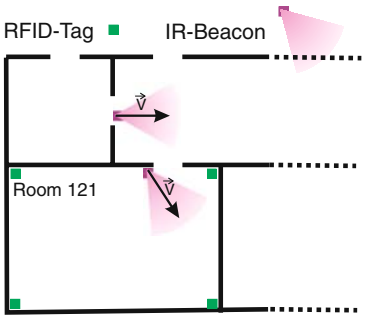


Fig. 10 Part of room plan with direction vectors of two IR beacons

5.2.4 Orientation Information Through Active RFID Tags

The transmission characteristic of active RFID tags is radial. Due to this radial transmission characteristic, the estimation of the direction is not as easy as with the infrared beacons. We estimate the direction as follows: (i) Store the starting position P_0 of the user in a variable *lastPosition*; (ii) With every new calculated position P_n, calculate the distance to *lastPosition*; (iii) If the distance is large enough (a few meters), calculate the direction vector $\mathbf{dir_{diff}} = lastPosition - P_n$ and store P_n in *lastPosition*. Repeat the steps (ii) and (iii) with every new calculated position.

5.2.5 Fusion of Orientation Information Through Bayesian Networks

The direction estimation with infrared beacons is rather accurate but the beacons are not always in reach of the user. The direction estimation through difference calculation is always possible (whether there are only RFID tags available or only infrared beacons or both) but it is also inaccurate. A combination of both techniques with a dynamic Bayesian network should give better and more stable results.

In this section we describe how the directional information from the infrared beacons and RFID tags can be combined.

As described in the introduction, we use a dynamic Bayesian network to fuse the direction data. The network itself is rather simple (see Fig. 11): It contains only three nodes and each node contains evidences for direction north, south, east, and west. The topmost node is the user direction node. This is the node that will contain the calculated (combined) direction after the roll-up and inference routines have been calculated.

The lower left node is the node for the infrared-based direction vector. As explained above, the estimated direction $\mathbf{dir_{IR}}$ is slightly inaccurate because the user can stand to the left or right of the main sending direction $\mathbf{v}$. This fact is encoded in the conditional probability table (CPT) of the infrared node: the probability that the user is heading in the estimated direction $\mathbf{dir_{IR}}$ is set to 0.9, the probability that she has a variation perpendicular to $\mathbf{dir_{IR}}$ is set to 0.045 (in both directions), and the probability that she is walking backward ($-\mathbf{dir_{IR}}$) is set to 0.01.

The right-hand node is the node for the direction that was calculated through the difference of two points ($\mathbf{dir_{diff}}$). Due to the fact that this direction is rather inaccurate, the CPT entries are as follows: 0.65 for heading exactly in direction $\mathbf{dir_{diff}}$, 0.15 for the variance perpendicular to $\mathbf{dir_{diff}}$, and 0.05 for walking backward.

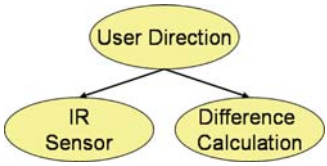


Fig. 11 Dynamic Bayesian network for direction fusion

5.2.6 Decomposition of a Direction Vector into Evidence Values

We need a way to represent an arbitrary direction vector as evidence values, so we can insert the estimated directions $\mathbf{dir_{IR}}$ and $\mathbf{dir_{diff}}$ in the respective nodes of the DBN. We do this by decomposing it in its x-component X and its y-component Y (see Fig. 12, left) and by making sure that the sum of the derived evidences is 1: If $Y > 0$ (this means the user has a north component in their direction), the evidence for north is

$$\text{e(north)} = \frac{Y^2}{X^2 + Y^2}$$

and the evidence for south e(south) is set to 0. If $Y < 0$ (the user has a south component in their direction), it is vice versa:

$$\text{e(north)} = 0 \text{ and e(south)} = \frac{Y^2}{X^2 + Y^2}.$$

The same principle applies for the x-component: If $X > 0$ (direction has east component), then

$$\text{e(east)} = \frac{X^2}{X^2 + Y^2} \text{ and e(west)} = 0.$$

If $X < 0$ (direction has west component), then

$$\text{e(west)} = \frac{X^2}{X^2 + Y^2} \text{ and e(east)} = 0.$$

Note that the sum of the evidences is always

$$\frac{X^2}{X^2 + Y^2} + \frac{Y^2}{X^2 + Y^2} = \frac{X^2 + Y^2}{X^2 + Y^2} = 1.$$

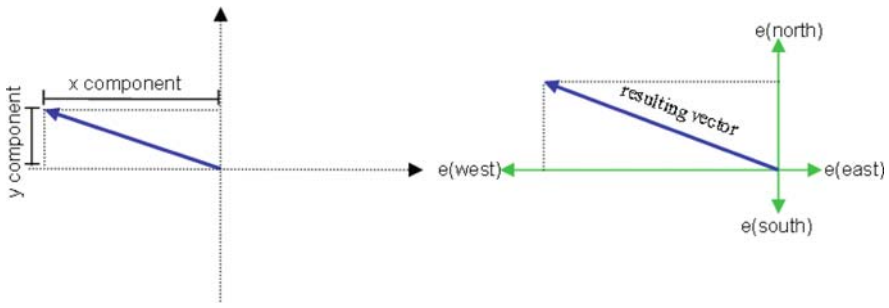


Fig. 12 Decomposition and composition of direction vector

5.2.7 Composition of a Direction Vector out of Evidence Values

The estimated directions $\mathbf{dir_{IR}}$ and $\mathbf{dir_{diff}}$ are decomposed as described above and the resulting evidences are inserted into the lower nodes of the DBN for each time slice, one time slice is added for every new measurement of the localization system. After performing the roll-up and the inference procedures of the DBN, the user direction node will contain evidences for north, south, east, and west components of the new direction. These evidence values must be combined to get a new direction vector. We do this by treating the components as a parallelogram of forces (as known from physics, see Fig. 12, right). The new direction vector $\mathbf{dir_{res}}$ consists of the x-component $X = \text{e(east)} - \text{e(west)}$ and the y-component $Y = \text{e(north)} - \text{e(south)}$ of the new calculated evidence values. The length of the calculated direction vector $\mathbf{dir_{res}}$ can be used as a confidence value, e.g., the longer the vector, the higher the confidence that the computed direction is correct.

5.2.8 Example Calculation

Figure 13 shows an example calculation. The top left node shows the result of the previous time slice, above it the composition of the calculated direction vector can be seen. Note that the evidence values for east and west are exactly the same, while the evidence for north is much bigger than for south. This causes the resulting vector point strictly to north. The two nodes below show the new evidences of the current measurement. The infrared direction has 100% evidence for east (the vector can be seen below the node), the difference direction node has the same evidence value for east and south.

The network on the right side of Fig. 13 shows the result after the inference routines have been carried out. The top node has strong evidence for east and still

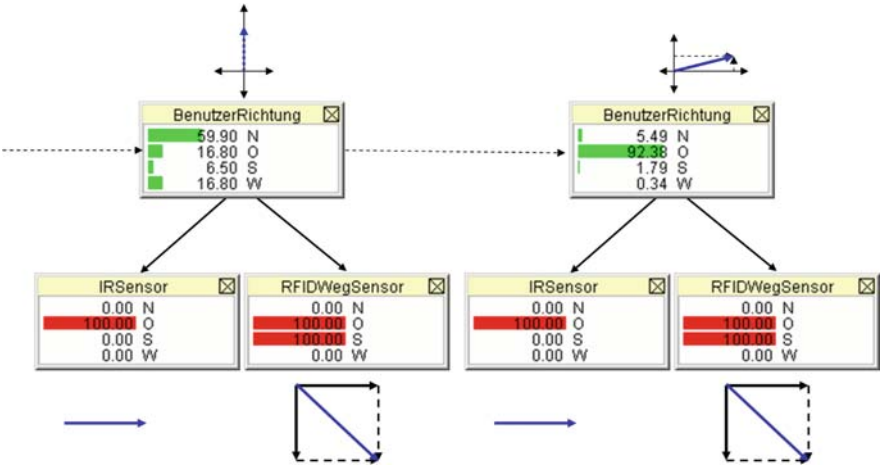


Fig. 13 Example calculation showing two time slices

little evidence for north. This is because both direction nodes give evidence for east but only the, perhaps inaccurate, difference direction node gives evidence for south. Because the DBN includes the previous calculated direction, the north component is still present in the new direction. This is the expected result of such a system, because it helps to smooth out fast jumping of the direction information and emphasizes the direction for which is has the most evidence.

After the detailed presentation of the core components in the preceding two chapters, we will now take a look at different applications whose services are based on the aforementioned components.

6 Implementing a Seamless, Proactive User Interface

In this section we pick up the idea to use animated 3D visualizations of indoor routes as navigational aid for pedestrians instead of static floor plans, which has been introduced earlier in [3] as a result of the SFB 378 project REAL. Visitors were able to choose a destination from a stationary information booth. Depending on their user profile and current stress level, such as time pressure, adaptive presentations have been rendered and were shown on a large screen. On their way, pedestrians were assisted by the mobile system IRREAL which iteratively presented new maps each time the user passed an infrared beacon. The concept of a seamless user interface for pedestrian navigation has been introduced in the BPN project, where a PDA connects three different situations and user interfaces: preparing a business trip at home, driving in the car, and finding the way from the car to the office in the street and in the building (see [14]).

We have reimplemented the functionality of the prior systems based on the current YAMAMOTO toolkit by using its routing and rendering engine, and the new positioning system LORIOT, so that the new systems, HYBNAVI and VISTO, are able to automatically generate arbitrary indoor routes across multiple levels. HYBNAVI is a tool for visualizing generated route descriptions. VISTO (VIdeos for SpaTial Orientation) [20] implements a seamless, proactive user interface based on the UBIDOO [23] task planner and the IPLAYBLUE framework [19] for ambient displays.

6.1 Hybrid Navigation Visualization

We developed a mobile navigation visualization system for pedestrians called HYBNAVI (HYBrid NAvigation VIsualization). In contrast to the BPN, all calculations, e.g., routes, route descriptions, and route visualization, are done on the mobile device of the user. HYBNAVI uses YAMAMOTO maps for these calculations and is not only able to navigate a user from one room to another in a multi story building, but also from a room in one building to another room in a different building. The system seamlessly switches from indoor to outdoor representations and dynamically re-calculates the route if the user deviates too far from the planned route.

Since HYBNAVI uses a VRML engine to render the YAMAMOTO map material, it can offer several 2D and 3D viewing perspectives. Based on psychological findings in [27] three perspectives which serve different purposes have been identified:

Immersed View. This is an egocentric perspective and shows the same view that the user currently has. Orientation and position of the camera are the orientation and position of the user. This view is best for navigation if the user is not required to gain a mental representation of the environment.

Tethered View. In this view the camera is positioned above and behind the user. Its orientation equals the orientation of the user. This view allows the user to gain more information about their surroundings, since it reduces the 'keyhole effect' of the immersed view.

Plan View. This is a 2D bird's eye view, where the camera is placed directly over the user's position. This is most favorable if the user wants to build a mental map of their surroundings.

Figure 14 shows each of the three views with the user standing at the same position. The use of YAMAMOTO map material also enables the system to show detailed objects, like soda machines or tables as seen in Fig. 15 (left and middle).

Fig. 14 Immersed view, tethered view, and plan view

Fig. 15 (*left, middle*) Detailed objects like soda dispensers and tables can act as landmarks. (*right*) HYBNAVI's view while navigating a user

These objects can serve as important landmarks for the user. Besides the rendering of the building interior, HYBNAVI also shows the planed route, directional arrows, a virtual compass, and an hourglass that visualizes the estimated walking time (see Fig. 15, right).

6.2 VISTO: Videos for Spatial Orientation

VISTO has been designed to proactively support arbitrary navigational tasks in indoor environments. The user interface is based on our own, web-based organizer called UBIDOO (UBIquitous To-DO Organizer).

In some public places you can already find information kiosk systems that present basic information about local services and maps for orientation. However, these systems are not frequently used even though they provide interesting information for the user. We believe this is because it takes too many steps of interaction until the desired information can be retrieved through the user interface. Typically, the user has to begin the dialogue with some basic choices, such as language, and navigate through many sub-menus. Often it feels easier to ask people or consult traditional maps, especially in situations of time pressure (e.g., if we arrive at the airport we would not like to experiment with some machine and try to figure out how it works). Yet we believe that public or ambient displays, if properly designed, have several advantages over personal devices. First, the diversity of mobile devices regarding their operating systems and technical capabilities makes it hard to implement assistance applications that run on any PDA or mobile phone. Second, all kinds of connectivity-related issues arise today, like WLAN access keys or different GSM frequency bands worldwide. Finally, mobile devices would require standardized indoor positioning services like LORIOT in order to provide navigational aid, which will not be available in the near future. Hence we use wall-mounted public displays whose position and orientation within the built environment can be statically modeled.

We have designed VISTO as a user-adaptive system which is able to automatically recognize the user's identity from their mobile device through the wide-spread Bluetooth wireless protocol, as the user passes an ambient display. The trick is that no data connection is required, so the users do not have to interact with their mobile device in order to set up anything except enabling Bluetooth. Once the user is identified, VISTO can retrieve information about the current tasks from the web-based UBIDOO organizer. UBIDOO applies Activity Theory in a sense that the user can hierarchically structure their tasks and relate them to a set of predefined activities, which are accessible through the UbisWorld ontology. Vice versa, the spatial location model of UbisWorld allows to associate places with typical activities. The combination of both relations, task-activity and task-location, yields in a powerful association between tasks and places. Thus VISTO can automatically sort out only relevant dates and tasks from the user's organizer and suggest a clear and useful set of destinations in the local area.

6.2.1 The Ubiquitous To-Do Organizer UBIDOO

With the advent of mobile computing devices, personal information management (PIM) tools have been implemented on PDAs and are now a common feature of mobile phones. Usually, the devices offer a calendar that allows to add events and set an alarm time for a reminder. The calendar is typically supplemented by a to-do list or notepad, where the user can add tasks that have no certain date. Whereas the conceptual model behind these PIM tools is well understood by both designers and users, it has obvious drawbacks that we want to address in this section.

Our ubiquitous task planner UBIDOO integrates the conceptual models of a diary and a to-do list into a hierarchical task tree, where each task can have multiple subtasks (see [10, 11]). UBIDOO is running on a web-server and can be accessed everywhere via the Internet. The user interface for desktop PCs is split in three frames, as shown in Fig. 16. On the left-hand side, the users' tasks and subtasks are shown in a foldable tree structure. Each task acts as a reminder to do a certain activity or action. The subtasks can either represent alternative choices, multiple goals, or fixed sequences of actions. Similar to a to-do list, the status of each task is graphically represented by a checkbox. By selecting a task from the tree on the left, a summary of its details (dates, location, and description) is shown and icons allow moving or deleting this task. In the frame to the right, various properties of the selected task can be edited which are arranged into six tabs. The bottom frame provides links to the user profile and a traditional calendar view. The user can also manually choose a location for an adapted task view ("here and now") that will be described later in more detail. For mobile devices, a reduced interface without frames and tabs is also available.

Usually, calendars in mobile devices offer an alarm function that reminds the user on events at a fixed time and date. Setting the proper time for an alarm requires the user to consider everything that needs to be done and prepared before the event, most importantly how to go there. Some events might require the user to go home first and dress accordingly or pickup some things, which takes additional time. The user has to plan and foresee the whole situation under uncertainty. However, often the situation changes in an unpredictable manner and we will not be where we have planned to be. Thus the alarm will remind us too early, unnecessarily distracting us from our work, or worse, remind us too late, and thus we cannot reach the event timely. Our ubiquitous task planner addresses this issue through an adaptive reminder, which continuously recalculates the best time for the alarm based on the current geographic distance between the user and the event. In addition, a general preparation time can be specified that will be considered by the reminder. Tasks can be localized by specifying a location from the spatial ontology in UbisWorld (see the "Place" attribute of a task in the "General" tab in Fig. 16). As the location of the user changes, the task planner updates the distance between the user and all tasks using route-planning web services. We use the eRoute service that is provided by PTV AG and return the estimated driving time between two locations. In the future, web services for public transportation and pedestrians could be easily added. The planner would then choose between car and bus based on the user's preferences.

Fig. 16 The Web-based user interface for the ubiquitous task planner UBIDOO

A further shortcoming of mobile organizers is their small display size. Browsing through calendar entries and large to-do lists is a cumbersome procedure. Hence we have implemented the "here-and-now" view in UBIDOO. It filters the user's tasks according to the current situation, which depends on time and location of the user, so that it displays only those tasks which can actually be worked on. The time horizon of this view is limited by the next binding date. The system performs a route calculation for each task to estimate the actual distance and time to get there. UBIDOO also considers the purpose of locations in terms of activities on a semantic level. If the user associates a task in the planner with activities from the ontology such as "shopping for electronics" instead of a certain store, the adaptive view automatically performs a search within UbisWorld's spatial elements and suggests the closest suitable location nearby for this task. Depending on the current time and location of the user, the system might suggest different places for the same task.

Figure 17 shows the same tasks as seen from two different locations. On the left image, we see the adapted view for Saarbrücken. The first task, swim and relax, is associated with the spatial purpose of waterworld and the planner suggests a place called Calypso in Saarbrücken that is located 7 km (or 11 min by car) from the current location of the user (at his office). A click on the R icon opens a route planner that provides a map and driving directions. The second task reminds the user to buy olives and milk. For the first item, a store called Saarbasar is set by the user as

Fig. 17 The location-adaptive task view as seen on a mobile device in Saarbrücken (*left*) and Munich (*right*)

location, whereas for the second item the general activity "shopping.food" has been specified so that the planner automatically suggests the Globus supermarket. The last task is a reminder to catch the flight from Zweibrücken; it takes 31 min to go there from the office. Now compare the times and distances with the adaptive view for Munich on the right. The olives do not appear, since the store is too far away. For the milk, a different store in Munich is suggested and also the waterworld has been substituted by a local place. The adaptive reminder for the airport will happen earlier, since the estimated traveling time is now more than 4 h.

6.2.2 The User Interface of VISTO

The VISTO user interface is designed to assist pedestrians in indoor environments and appears on wall-mounted public situated displays, which are spread across the building at decision points. The screen layout is split into three areas as shown in Fig. 18. On the top border, a status bar shows the name of the location and the registered activities which are used to search and filter the UBIDOO task tree for. The status bar also includes a personalized icon for each user that has been currently recognized by the Bluetooth scanner. The user icons are continuously updated after each scan (each scan takes approximately 15 s; single misses of a previously recognized device are ignored to stabilize the displayed list).

On the left-hand side the activities of the recognized users are shown as blue-labeled tabs. They show a list of tasks (actions), according to the hierarchic task tree structure in UBIDOO, which remind the user of his objectives depending on

Fig. 18 VISTO is designed to run on public displays and presents an adaptive list of goals (*left*) and animated walking directions (*right*)

the current spatial context. In the given example, one tab represents a shopping list of products, another tab a list of persons to visit. In the context of VISTO, each list item represents a navigational goal and provides some basic information to the user about the walking direction and distance (relative to the current display's orientation and location, which is known to the system through the YAMAMOTO environment model). The tasks are sorted by increasing distance, so that the closest goal appears on top of the list. If the display affords interaction by means of a touchscreen, keyboard, or mouse, the user can activate any task from the list by clicking on its arrow symbol. There can be only one active task per user. On the right-hand side, video frames present detailed walking directions to each active goal as a 3D animation. Each animation shows the route from the egocentric perspective of the user. The movie starts at the current location and the avatar is facing the respective display within the virtual environment model. The avatar follows the path until the destination or another display is reached. The movie loops until the user leaves the Bluetooth range of the display. If the user encounters a new display along the route, the playback of the next movie sequence is automatically triggered without any interaction. Finally, if the user has reached their destination, the current task will be deactivated. In summary, VISTO is a user-adaptive system that adapts its user interface according to the current activity of the user and proactively assists them in their wayfinding tasks.

7 Summary

In the chapter at hand, we have elaborated on the development of a map modeling toolkit for buildings in 3D as well as on a new indoor positioning system. The toolkit meets the demands of providing map-like material for pedestrian indoor navigation and allows for a quick and easy editing of indoor map material. In addition, it implements a route finding service for buildings. In parallel, we have implemented a new indoor positioning system. The positioning system is able to combine active RFID tags, infrared beacons, and GPS to calculate the users' position following the "always best positioned" paradigm in a resource-adaptive manner. On the basis of those core components, we introduced different visualization techniques for route descriptions using personal digital assistants or public displays available in the environment. The whole system, as explained in the chapter present, is able to seamlessly combine indoor and outdoor navigation tasks in such a way that the changeover between different localisation techniques is barely noticeable for the user. The results of the project led to several new external cooperations, e.g., Fraunhofer IESE Kaiserslautern and LMU Munich, which have been granted a research licenses for LORIOT and YAMAMOTO. Furthermore, they led to the foundation of a new spin-off company, schwartz&stahl indoor navigation solutions GbR (http://www.indoornavi.com) in order to bring the results of the project to market.

Within the project, we intensified our cooperation with the research group of the project AERA of the Collaborative Research Center 378, Resource-adaptive Cognitive Processes. During the last period, both projects, AERA and RENA, have been closely cooperating in the design of an experimental paradigm to investigate spatial navigation and wayfinding with the aim to optimize spatial instructions during the navigation task. The results of a series of experiments are summarized and reported by [28] in this volume and in detail in [1, 16, 17].

Acknowledgments This research has been supported by the German Science Foundation (DFG) in its Transfer Unit 53, Project RENA: Resource-Adaptive Navigation.

References

1. Aslan, I., Schwalm, M., Baus, J., Krüger, A., Schwartz, T. Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems. In: Proceedings of the 8th international Conference on Human Computer Interaction with Mobile Devices and Services (Mobile HCI 2006) (pp. 105–108). ACM Press (2006).
2. Bahl, P., Padmanabhan, V. RADAR: An in-building RF-based user location and tracking system. IEEE INFOCOM, 2:775–784 (2000).
3. Baus, J. Ressourcenadaptierende Hybride Personennavigation. DISKI 268. Berlin: Akademische Verlagsgesellschaft Aka GmbH (2003).
4. Baus, J., Krüger, A., Stahl, C. Resource-adaptive personal navigation. In: O. Stock, M. Zacanaro (Eds.), Mulitmodal Intelligent Information Presentation (pp. 71–93). New York: Springer (2005).

5. Beer, T. Ein System zur Gefahrenmeldung basierend auf Methoden des Wissensmanagements und des Semantic Web. Master's thesis, Universität des Saarlandes (2004).
6. Brandherm, B., Schwartz, T. Geo referenced dynamic Bayesian Networks for user positioning on mobile systems. In T. Strang, C. Linnhoff-Popien (Eds.), Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, Lecture Notes in Computer Science (vol. 3479/2005, pp. 223–234). Munich, Germany: Springer-Verlag Berlin Heidelberg (2005).
7. Department of Defense: Department of Defense World Geodetic System 1984. Tech. Rep. Third Edition, National Imagery and Mapping Agency (2000).
8. Ehrl, J. Prototypische Entwicklung eines kontextsensitiven, regelbasierten Dienstes zur lokalen Gefahrenwarnung im automobilen Umfeld mit Technologien des Semantic Web. Master's thesis, Universität des Saarlandes (2005).
9. Ekahau, Inc. Ekahau Positioning Engine. `http://www.ekahau.com/products/positioningengine/` (2004).
10. Fickert, O. Ein Ubiquitärer Aufgabenplaner für Benutzergruppen. Tech. Rep., Unversität des Saarlandes (2005).
11. Fickert, O. Ein Ortsbezogener Termin- und Aufgabenplaner mit Routenwissen. Master's thesis, Unversität des Saarlandes (2007).
12. Gholamsaghaee, E. Ein System zur Positionsbestimmung in Parkäusern mittels aktiven RFID-Tags. Master's thesis, Universität des Saarlandes (2007).
13. Johanson, B., Fox, A. The event heap: A coordination infrastructure for interactive workspaces. In: Proceedings of the Fourth IEEE Workshop on Mobile Computing Systems and Applications (pp. 83–93). Callicoon, New York (2002).
14. Krüger, A., Butz, A., Müller, C., Wasinger, R., Steinberg, K., Dirschl, A. The connected user interface: Realizing a personal situated navigation service. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI 2004) (pp. 161–168). ACM Press, New York (2004).
15. Müller, J. Beanspruchungsschätzung im Automobil mit Bayes'schen Netzen. Master's thesis, Unversität des Saarlandes (2005).
16. Münzer, S., Stahl, C. Providing individual route instructions for indoor wayfinding in complex, multi-level buildings. In: GI-Days 2007, ifgi prints, Münster (2007).
17. Münzer, S., Zimmer, H.D., Maximilian Schwalm Jörg Baus, I.A. Computer assisted navigation and the acquisition of route and survey knowledge. Journal of Environmental Psychology 26:300–308 (2007).
18. Mußler, O. Verfahren zur Positionsbestimmung in Parkhäusern. Master's thesis, Universität des Saarlandes (2004).
19. Schöttle, F. IPlay Blue: Interaktive Displays mit Bluetoothbasierter Benutzererkennung. Tech. Rep., Unversität des Saarlandes (2006).
20. Schöttle, F. Automatisches Generieren von graphischen Wegbeschreibungen für Fußgänger. Master's thesis, Unversität des Saarlandes (2007).
21. Schwartz, T., Brandherm, B., Heckmann, D. Calculation of the user-direction in an always best positioned mobile localization system. In: Proceedings of the International Workshop on Artificial Intelligence in Mobile Systems (AIMS) (2005).
22. Stahl, C., Haupert, J. Simulating and evaluating public situated displays in virtual environment models. In T. Pederson, H. Pinto, M. Schmitz, C. Stahl, L. Terrenghi (Eds.), International Workshop on Modelling and Designing User Assistance in Intelligent Environments (MODIE 2006) (pp. 32–35) (2006).
23. Stahl, C., Heckmann, D., Schwartz, T., Fickert, O. Here and now: A user-adaptive and location-aware task planner. In S. Berkovsky, K. Cheverst, P. Dolog, D. Heckmann, T. Kuflik, J. Picault, P. Mylonas, J. Vassileva (Eds.), Proceedings of the International Workshop on Ubiquitous and Decentralized User Modeling (UbiDeUM'2007) (pp. 52–63) (2007).
24. Wahlster, W., Feld, M., Gebhard, P., Heckmann, D., Jung, R., Kruppa, M., Schmitz, M., Spassova, L., Wasinger, R.: The Shopping Experience of Tomorrow: Human-Centered and Resource-Adaptive. In: M. Crocker, J. Siekmann (Eds.) Resource Adaptive Cognitive Processes (this volume), Cognitive Technologies Series. Springer

25. Ward, A., Jones, A., Hopper, A. A new location technique for the active office. IEEE Personal Communications 4(5):42–47 (1997).
26. Wasinger, R., Stahl, C., Krüger, A. M3I in a Pedestrian navigation & exploration system. In L. Chittaro (Ed.), Proceedings of the 5th International Symposium on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI) (pp. 481–485). Udine, Italy: Springer (2003).
27. Wickens, C., Hollands, J. Engineering Psychology and Human Performance. Upper Saddle River, NJ: Prentice Hall (2000).
28. Zimmer, H.D., Münzer, S., Baus, J.: From Resource-adaptive Navigation Assistance to Augmented Cognition. In: M. Crocker, J. Siekmann (Eds.) Resource Adaptive Cognitive Processes (this volume), Cognitive Technologies Series. Springer

Linguistic Processing in a Mathematics Tutoring System: Cooperative Input Interpretation and Dialogue Modelling

Magdalena Wolska, Mark Buckley, Helmut Horacek, Ivana Kruijff-Korbayová, and Manfred Pinkal

1 Introduction

Formal domains, such as mathematics, require exact language to communicate the intended content. Special symbolic notations are used to express the semantics precisely, compactly, and unambiguously. Mathematical textbooks offer plenty of examples of concise, accurate presentations. This effective communication is enabled by interleaved use of formulas and natural language. Since natural language interaction has been shown to be an important factor in the efficiency of human tutoring [29], it would be desirable to enhance interaction with Intelligent Tutoring Systems for mathematics by allowing elements of mixed language, combining the exactness of formal expressions with natural language flexibility. Natural language itself is in a sense inherently inappropriate for formal domains as it introduces imprecision and ambiguity into the formalised environment. However, in interaction between humans identified ambiguities are contextually resolved, and tutorial systems should also offer some degree of cooperative ambiguity resolution.

A minimal approach to dialogue-based tutoring would only offer evaluations of the correctness of the student's input. However, there is evidence that the effectiveness of human tutoring relies on the tutor monitoring the student's state of knowledge, identifying possible misconceptions and applying pedagogical strategies to guide the student to a solution to the exercise [16]. In order to support this kind of interaction it is necessary to maintain a model of the current state of the interaction. Information state in dialogue systems has been modelled as a structure including information about the linguistic and domain-level analyses of the contribution, as well as the context in which the contribution has been uttered, that is, information arising from the contributions from the previous discourse [36]. However, given the tutorial setting, the model has to be updated in a manner which accounts for the phenomena found in tutorial dialogue. The model of the dialogue state can be used

M. Wolska (✉)
Department of Computational Linguistics & Phonetics, Saarland University, 66123 Saarbrücken, Germany
e-mail: magda@coli.uni-sb.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_12,

to make inferences, for example, about (1) which facts the student assumes to be known, (2) which false statements the student has mistakenly come to believe, and (3) which true statements the student erroneously believes to be false. Such a model can feed into a module that monitors the student's performance [27, 28], which in turn forms the context in which a pedagogically appropriate action can be chosen.

Our research on language and dialogue aspects of mathematics tutoring was carried out within the SFB 378 project DIALOG[1] [7] which aimed at partially automating intelligent tutoring of mathematical proofs. In the proof tutoring scenario, as studied in the project, a student solves exercises by communicating proof steps, possibly expressed in natural language, to a tutorial system. A number of foundational research challenges have been identified that are crucial to realising intelligent computer-supported natural language proof tutoring, including linguistic and mathematical analysis of students' input, representation and maintenance of the proof state, proof step evaluation, dialogue modelling, and pedagogical issues. These aspects are introduced in "Resource-Bounded Modelling and Analysis of Human-Level Interactive proofs" by Benzmuller et al., in this volume, which treats modelling and analysis of interactive proofs in detail.

In order to investigate the characteristics of the language and the dialogue structure in the context of computer-based tutoring of mathematics, we conducted two experiments in which we collected corpora of simulated tutorial dialogues in mathematical theorem proving. The analysis of the data allowed us to identify the prominent linguistic phenomena at both the utterance and the dialogue level. In this chapter, we present research which addresses two of the general issues relevant to natural language mathematics tutoring pointed out by Benzmuller et al., namely techniques of input interpretation and dialogue modelling. We organise our presentation of language processing by associating the phenomena with one of three strata, depending on the techniques required in interpretation: (1) basic architecture for analysis, (2) accounting for mixed language and contextual ambiguity resolution, and (3) error-tolerant analysis. These techniques are characterised by a tight interaction between natural language parsing and mathematical formula interpretation, extended representation of mathematical knowledge, and error-tolerant interpretation of formulas. At the dialogue level, we show how the knowledge stored in the dialogue state can be used to support both cooperative interpretation and the choice of system action, in particular in the case of knowledge misalignment.

The paper is organised as follows: First we present the project setting in which our work is embedded. In Sect. 3 we characterise the language of proofs in the collected corpora. Section 4 discusses requirements and describes interpretation techniques in the basic processing architecture. Section 5 elaborates on domain and context-specific extensions required by more complex phenomena along the three strata mentioned above. In Sect. 6 we discuss dialogue modelling in a tutorial

[1] The DIALOG project was a collaboration between the Computer Science and Computational Linguistics departments of Saarland University within the Collaborative Research Center on *Resource-Adaptive Cognitive Processes*, SFB 378 (`http://www.coli.uni-sb.de/sfb378`).

scenario, present an integrated model of dialogue state and show its use in modelling common ground. Finally, we present related work and summarise our contribution.

2 Research Setting

The work described in this chapter fulfilled the basic purpose of providing required components for a dialogue-based tutorial system, which formed one of the DIALOG project's major goals. We will sketch the overall architecture of the system in the main part of this section. Beyond the immediate purpose of contributing to the system prototype, we consider processing of interleaved symbolic and informal natural language discourse, as it occurs in mathematics tutoring, to be an interesting research objective in itself, dealing with a challenging variant of multimodal interaction. Moreover, methods of flexible linguistic processing presented in this chapter support the system's ability to adapt to the communicative needs and the mathematical reasoning skills of students at different levels of expertise, accepting all varieties of student input, ranging from mathematical formulas alone to colloquial natural language including vagueness, ambiguity, and ill-formed expressions. Thus, this chapter highlights one facet of the general research theme of resource-adaptive cognitive processing, the objective of the SFB 378 of which the project was a part. We now briefly introduce the components of the system architecture which is schematically presented in Fig. 1.

Dialogue Manager: The system design is centred around a dialogue manager that obtains information from a number of specialised modules in order to maintain and update the representation of the dialogue context and to drive the dialogue forward by choosing the system's responses. Dialogue management and the treatment of certain tutoring-specific dialogue phenomena are further discussed in Sect. 6.

Natural Language Understanding (NLU): The first step of the processing pipeline is the interpretation of the input utterances. The task of the input analysis module is two-fold. First, it is to construct a representation of the utterance's linguistic meaning. Second, it is to identify and separate within the utterance: (i) the parts which constitute meta-communication with the tutor (e.g. "Ich habe die Aufgaben-

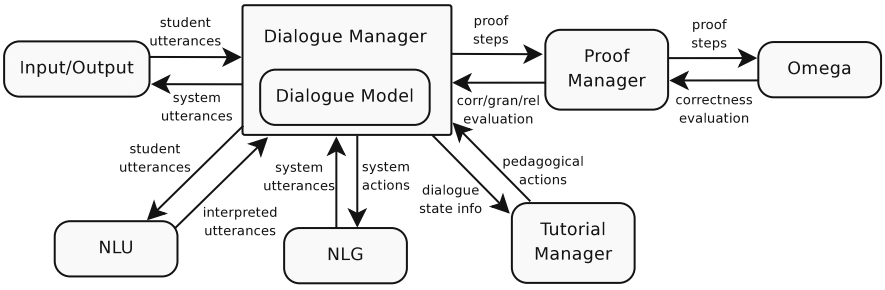


Fig. 1 Architecture of the System

stellung nicht verstanden." ["I don't understand what the task is."]) that are not to be processed by the domain reasoner, and (ii) parts which convey domain knowledge, *domain contributions*, that should be verified by a domain reasoner. Language phenomena specific to learner mathematical discourse are discussed in Sect. 3. Input analysis is discussed in Sect. 4 and 5.

Proof Manager (PM): The main task of the proof manager is to build and maintain a representation of the proof(s) constructed by the student. The PM communicates with external components to obtain evaluations of the proposed proof steps with respect to their *correctness*, *granularity*, and *relevance*.

Mathematical Assistant System: The core domain reasoning component is an automated theorem proving system. The system accepts *underspecified* proof steps represented at the *assertion level* [40] and checks the appropriateness of proof steps with respect to their *correctness* in a valid proof by attempting to incorporate the alternative proof step interpretations into the proof state representation constructed so far. The system also provides information needed in the evaluation of granularity and relevance aspects of proof steps. The mathematical assistant system we used in the project is the ΩMEGA system "ΩMEGA: Resource-Adaptive Processes in Automated Reasoning system" by Auterier et al., this volume. The use of ΩMEGA in the evaluation of proof steps' granularity and relevance is discussed in "Resource-Bounded Modelling and Analysis of Human-Level Interactive proofs" by Benzmuller et al., in this volume.

Tutorial Manager: The tutorial manager stores *pedagogical knowledge*, the main component of which is a set of teaching strategies, in particular knowledge on hinting strategies for mathematical proofs. The categorisation of hints and production of hints in mathematical theorem proving are presented in detail in [37, 38].

3 The Language of Informal Proofs

In formal domains, such as mathematics, solutions to problems found in course books differ from informal presentations in student–tutor interactions. While there are plenty of sources to study text book mathematics, there is hardly any material on *dialogues* in this domain. In order to investigate the use of language in this setting and to identify requirements on automating discourse interpretation we conducted two experiments in a simulated (*Wizard-of-Oz*) setting. The details of the experiments are presented in [4–6, 42, 43] and summarised in "Resource-Bounded Modelling and Analysis of Human-Level Interactive proofs" by Benzmuller et al., in this volume.

The second experiment was a pilot study that sought to see whether the mode of presentation of the study-material influences the style of interaction with the system. The subjects were divided into two groups and were presented with either verbose study-material (using a mixture of natural language and formulas) or formal presentation (using mainly formulas). Preliminary results of the analysis of the

Table 1 Overview of the corpora

	#subjects	#turns	#S turns	#T turns
Corpus-I	22	775	332	443
Corpus-II	37	1917	937	980
FM-group	20	974	474	500
VM-group	17	943	463	480

second corpus with respect to language production in the two conditions have been presented in [43]. Table 1 presents a quantitative overview of the two corpora.[2]

Below, we present an overview of the prominent language phenomena observed in our corpora to illustrate the characteristics of the language used in this setting. We focus on the mixed language combining mathematical expressions and natural language, imprecision and ambiguity of the natural language, referring expressions, and ill-formed and invalid mathematical expressions. For clarity of presentation of the corpus examples, we give only the English translations of the German data, provided the translation preserves the phenomena of interest. The letters K and P stand for set complement and powerset, respectively.

Mathematical formulas and mixed language The formal language of mathematics consists of symbolic expressions. (1) is an example of such an expression from the domain of naive set theory. One of the most prominent characteristics of the language of informal mathematics is that symbolic mathematical expressions (in a learning setting often semi-formal or ill-formed) and natural language are intermixed as in (2) through (5). In particular, formulas (or parts thereof) may lie within the scope of quantifiers or negation expressed in natural language as in (4). Utterances may contain any combination of symbolic and worded content: from formal alone, as in (1), mixed symbolic and natural language expressions, as in (2) through (4), up to entirely worded natural language sentences, as in (5).

(1) $A \cap B \in P(A \cap B)$
(2) $A \cap B$ is $\in$ of $C \cup (A \cap B)$
(3) According to DeMorgan-1 $K(A \cup B)$ equals $K(A) \cap K(B)$
(4) B contains no $x \in A$
(5) A contains no elements that are also in B

Imprecise language Instead of constructing formal mathematical expressions, students tend to use imprecise and informal descriptions introducing ambiguity. The corpora contain multiple examples of formulations that are imprecise or ambiguous from the structural or lexical point of view. Structural ambiguities may occur at the formula level, utterance structure level, and proof-structure level. Lexical ambiguities occur where natural language words are used in place of mathematical

[2] Labels "FM-group" and "VM-group" refer, respectively, to the groups presented with formal and verbose presentation of the study-material. The columns "#S turns" and "#T turns" summarise the number of student and tutor turns.

terminology to name mathematical concepts. Examples of imprecise statements are presented below.

(6) $x \in B$ and therefore $x \subseteq K(B)$ and $x \subseteq K(A)$ given the assumption
(7) If the union of A and C is equal to intersection of B union C, then all A and B must be contained in C
(8) then A and B are entirely different, have no common elements
(9) by deMorgan-Rule-2 it holds that $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$
(10) deMorgan-Rule-2

An ambiguous coordination introduces a structural ambiguity in (6).[3] In (7), as well as (4) above, the words "contain" and "be in" are ambiguous with respect to the intended meaning of the *Containment* relation they evoke. Depending on context, *Containment* may be plausibly interpreted as, among others, *Structural composition* or *Inclusion*.[4] The latter is further ambiguous between the relations of (STRICT) SUPER-/SUBSET or ELEMENT in the context of naive set theory. In (4) both readings are possible.[5] (8) is an example of an informal worded description of the empty intersection of two sets. Finally, (9) and (10) exemplify ambiguities at the proof structure level. Both were uttered by students as first dialogue turns in response to the following problem: $K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))$. In (9), it is not clear whether the provided formula is to be interpreted as an instantiation of the DeMorgan rule or as a consequent of the formula to be proved. In (10), the student does not indicate how DeMorgan rule is to be applied.

Referring Aside from imprecise formulations, we observed the use of domain-specific referring expressions whose interpretation requires a metonymic extension. For example, "the left hand side" in (11) refers to a part of an equation. Other expressions of the same nature include "the parenthesis", "the left parenthesis", and "the inner parenthesis", as in (12), that are used to refer to bracketed sub-expressions. Similarly, "complement", in (13), does not refer to the operator per se, but rather to the expression in the operator's scope, that is, an expression in which "complement" is the top-level operator. Further examples of referring expressions in our corpora can be found in [44].

(11) Then for the left hand side it holds that $C \cup (A \cap B) = (A \cup C) \cap (B \cup C)$, the term $A \cap B$ is already there, and so an element of it
(12) *T1:* Bitte zeigen Sie: $A \cap B \in P((A \cup C) \cap (B \cup C))$!

[3] The alternative readings are: "($x \in B$ and therefore $x \subseteq K(B)$) and ($x \subseteq K(A)$ given the assumption)" and "([$x \in B$] and therefore [$x \subseteq K(B)$ and $x \subseteq K(A)$] [given the assumption])"
[4] In the following we use SMALL CAPS to refer to domain specific relations and *italics* to refer to general concepts.
[5] The *Structural composition* reading may be intended if in the previous context there is an assignment of B to a formula in which $x \in A$ is a sub-expression. The ELEMENT-*Inclusion* reading might be possible if B is a set whose elements are formulas.

S1: Distributivität von Vereinigung über Durchschnitt: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ Hier dann also: $C \cup (A \cap B) = (A \cup C) \cap (B \cup C)$ Dies für die innere Klammer
T1: Prove: $A \cap B \in P((A \cup C) \cap (B \cup C))$!
S1: Distributivity of union over intersection: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ Here: $C \cup (A \cap B) = (A \cup C) \cap (B \cup C)$ This for the inner parenthesis
(13) de morgan rule 2 applied to both complements

Ill-formed mathematical expressions The symbolic expressions are often ill-formed. In (14), the bracketing required for operators of the same precedence is missing, which makes the formula ambiguous. In (15) parentheses delimiting the scope of the powerset operator are missing on the right-hand side. (16) is an example of a typographical error, where an operator symbol p is used in place of the identifier b. In (17), arguments of the main operator, $\notin$, are of the wrong type.

(14) $A \cup D = A \cup B \cap C \cup D$
(15) $P((A \cup C) \cap (B \cup C)) = PC \cup (A \cap B)$
(16) $(p \cap a) \in P(a \cap b)$
(17) $(x \in b) \notin A$

Semantic and pragmatic errors in mathematical statements Finally, some student statements do not convey the expected semantics. In (18), the student makes a statement about set equality, but in the given proof context, a weaker assertion about the set inclusion ($\subseteq$) is needed. (19) is an example of commonly confused ELEMENT and SUBSET relations.

(18) $P((A \cap B) \cup C) = P(A \cap B) \cup P(C)$
(19) If $A \subseteq K(B)$ then $A \notin B$

Our goal is to design an input interpretation strategy for informal discourse on mathematical proofs in which symbolic and natural language expressions may be freely intermixed. Considering the language phenomena illustrated above, we identify four sub-tasks in the input understanding process: (i) identification and analysis of the mathematical expressions, (ii) parsing the mixed language utterance, (iii) interpretation of ambiguous natural language formulations, and (iv) constructing a logical representation of the propositional content of the proof-step. Additionally, following a human-tutor strategy, our aim is to attempt to cooperatively recover the intended meaning of ill-formed, sloppy, imprecise, or incomplete specifications. By doing this, we allow the student to concentrate on the higher problem solving goal, rather than engaging him/her in tedious clarification sub-dialogues.[6]

In the following sections, we first present the core interpretation architecture, and second, specific extensions to the analysis process required to account for certain amount of imprecision and incompleteness in students' specifications of proof-steps. We identify the knowledge sources used in the course of input analysis and discuss

[6] The strategy, however, would change, for instance, when attempting to teach correct terminology, an issue we do not address in this chapter.

their role in the interpretation. Furthermore, we present means of *cooperative* interpretation by pointing at ways of recovering from students' low-level errors.

4 Baseline Processing

The baseline processing architecture involves basic functionality which is required to interpret simple cases of mixed language statements. Below, we briefly outline the processing steps and present an example analysis.[7]

Mathematical expression parsing Analysis of mathematical expressions is a preprocessing stage of syntactic and semantic parsing. The mathematical expression recogniser and parser use knowledge of operators and identifiers relevant in the domain and the given problem to identify the type of the expression and specific aspects of its structure. For the purpose of syntactic and semantic parsing, the original utterance is re-interpreted in terms of an abstract representation of the formal content: mathematical expressions are assigned symbolic tokens representing their types. For example, the expressions $A \cup B$ and $K(A) \cap K(B)$ in (3) are assigned the tokens `TERM`. The utterance parser operates on these symbolic tokens.

Syntactic and semantic analysis We adopt a strategy of deep syntactic and semantic linguistic analysis in order to achieve a systematic and consistent account of the mixed-language discourse. The task of the parser is to deliver representations of the *linguistic meaning* of sentences and syntactically well-formed fragments.

As the linguistic meaning representations, we adopt the relational dependency structures of the *tectogrammatical level* from [34]. The central unit of a clause is the head verb that, in its valency frame, specifies the *tectogrammatical relations* of its dependents (also referred to as *participants/modifications*).

To build the linguistic meaning representations, we use a lexically based Combinatory Categorial Grammar parser, OpenCCG.[8] Our motivation for using this grammar framework is two-fold: first, it is known for its account of coordination phenomena, widely present in mathematical discourse, and second, mathematical expressions, represented as their types, lend themselves to a perspicuous categorial treatment as follows: In the course of parsing, we treat symbolic tokens that represent mathematical expressions on a par with lexical units. The parser's lexicon encodes "generic" lexical entries for each mathematical expression type (represented by its token, e.g. `TERM`, `FORMULA`), together with information on the syntactic categories the expression may take (e.g. the category of a noun phrase, `np`, for `TERM` and the category of a sentence, `s`, for `FORMULA`). The choice of syntactic categories for the mathematical expression tokens was guided by a systematic study of the syntactic contexts in which mathematical expressions are used in our tutorial dialogue corpus and mathematical textbooks. More details on parsing issues are presented in [42].

---

[7] We omit here the obvious pre-processing such as sentence- and word-tokenisation.

[8] `http://openccg.sourceforge.net`

Step-wise interpretation The linguistic meaning representations built at the parsing stage do not contain any information as to how the utterance is interpreted in the context of the given domain. Interpretation is a step-wise procedure in which predicates and relations of the linguistic meaning representations are gradually assigned domain-specific semantics.

First, semantemes are mapped to concepts through a semantic lexicon. The mapping serves either to recast the tectogrammatical frames in terms of conceptual predicates and roles, or provides procedural recipes to compute lexical meanings. For example, the `Norm` relation introduces a concept of a *Rule* and the `Norm` dependent is the *Rule* according to which the head proposition holds. Second, a domain ontology is consulted to find domain-specific interpretations corresponding to the concepts from the semantic lexicon. The role of the ontology is to provide domain-specific information of imprecise linguistic formulations. More details on the motivation for the intermediate representation and its structure have been presented in [21].

Figure 2 shows the interpretation of the utterance (3): "According to DeMorgan-1 $K(A \cup B)$ equals $K(A) \cap K(B)$". After pre-processing, the utterance is converted into a form that abstracts from the mathematical expressions: "According to DeMorgan-1 `TERM` equals `TERM`". The linguistic meaning representation is presented on the left and consists of the German copula, "ist", with the symbolic meaning **equals**, as the head of the structure and three dependents in the relations of `Actor`, `Norm`, and `Patient`. To the right, the step-wise domain meaning assignment is shown: First, based on the semantic lexicon, a concept of *Equivalence* is assigned to "ist", with the `Actor` and `Patient` dependents as relata (Arg-1 and Arg-2), and the `Norm` dependent is interpreted as *Rule*. Next, *Equivalence*, in the context of set theory `TERM`s, is interpreted as *equality* (=), and the *Rule*, in the context of theorem proving, as *Justification* in a proof-step.

Output Structure The output of the input analysis component is the underspecified representation of the proof step used by the domain reasoner [1]. For the example in Fig. 2 it is "*Assertion* $K(A \cup B) = K(A) \cap K(B)$ *Justification* DeMorgan-1". Each output is moreover labelled with information relevant for the Dialogue Manager. For instance, the fact that the given output structure represents a proof step is marked as `domain contribution`, the fact that the utterance is linguistically marked as containing information assumed to be known (by a modal particle "doch" for instance), i.e. it is informationally redundant is marked as IRU. In Sect. 6 we

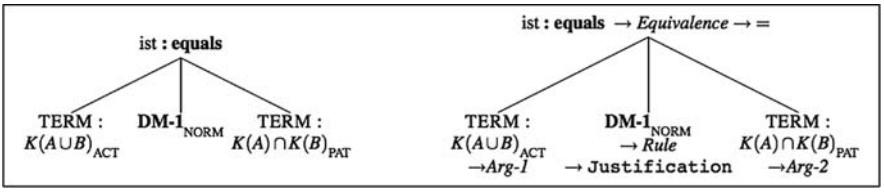


Fig. 2 Interpretation of utterance (3); DM-1 stands for DeMorgan-1

show how this marking is used by the Dialogue Manager to update the dialogue model.

5 Aspects of Cooperative Interpretation

We have developed methods for cooperative interpretation of the student utterances using automatic error correction procedures. By "cooperative" we mean that our goal is (i) to use domain reasoning to attempt to recover from low-level errors and (ii) to accommodate the appropriate readings of informal and ambiguous statements when there is no need for immediate clarification. In the latter case, we make assumptions on the intended semantics by finding plausibly intended, formally correct interpretations. In this section, we discuss some of the issues involved in pursuing this strategy.

The basic processing presented in the preceding section covers simple utterances, ((1), (3), (6)),[9] but does not account for more complex ones. Below, we identify areas that require extensions and exemplify the relevant phenomena:

- *parsing issues:* incomplete mathematical expressions used as short-hand for natural language, as in (2), scope phenomena involving parts of formula, (4), use of spoken-language syntax to verbalise mathematical expressions, (7);
- *lexical semantics and domain modelling:* imprecision and ambiguities introduced by natural language, (5) and (7), informal natural language formulations, (8);
- *reference resolution:* domain-specific metonymic references, (11) and (13);
- *disambiguation:* relevance of domain reasoning in resolving ambiguities and linguistically non-ambiguous, but task-level ambiguous proof-steps, (9) and (10).

A manually constructed intermediate domain model, the domain ontology, plays a crucial role in extending the basic functionality in the areas identified above. It is a hierarchically organised representation of objects and relations relevant in the domain, together with their properties, that allows type checking and reasoning about relations between objects. The core features of the model relevant in extending the interpretation coverage include representation of (i) *imprecise and general concepts* that have a meaning independent of the mathematical domain, but need to be interpreted in terms of their domain-specific readings (in the relations ontology), (ii) *typographical properties* of mathematical objects, including substructure delimiters, linear orderings, delimited substructures (in the objects ontology). The purpose of the former is to allow us to interpret ambiguous relations. The latter provides access to substructures of mathematical expressions as objects for anaphoric reference. In particular, given this knowledge, we can identify types of the subcomponents and reason about them.

[9] The example numbers in this section refer to example utterances on pp. 271 and 272.

Below, we elaborate on the extensions to the basic processing and point out the role of the following knowledge sources in extending the capacity of the interpretation component: the formula parser's knowledge of the structure of mathematical expressions, the semantic lexicon, the domain ontology, the domain reasoner, and the pedagogical resources.

5.1 Parsing

The mixed language mode consisting of natural language interleaved with formal expressions calls for dedicated extensions to parsing procedures. The most important information is that of the type of the mathematical expression embedded within natural language text. Below, we outline the specific necessary extensions.

Incomplete mathematical expressions Both the mathematical expression parser and the natural language parser are adapted to support incomplete mathematical expressions and their interactions with the surrounding natural language text. In particular, the formula tagger and parser recover information on incomplete expressions using knowledge of syntax and semantics of formal expressions in the domain. For example, in (2), the operator $\in$ is recognised and, based on domain knowledge, identified as requiring two arguments: one of type *inhabitant* and the other *set*. Accordingly, it is assigned a symbolic type `0_FORMULA_0`, where `0` indicates the argument missing in the left and the right context. Furthermore, a lexical entry `0_FORMULA_0` with the syntactic category $\mathtt{s/pp}_{lex:von}\mathtt{\backslash n}$ is included in the lexicon of the syntactic parser. Other kinds of incomplete mathematical expressions and their types are treated in a similar way by identifying their incomplete type, used as token during parsing, and introducing a corresponding entry in the parser's lexicon.

Interactions with natural language To account for interactions between mathematical expressions and the surrounding natural language text, as in (4), we identify structural parts of mathematical expressions that may lie within the scope of a natural language expression. Similarly to above, we use domain knowledge about the types of mathematical expressions to identify the substructure and assign incomplete types to the relevant substructures that may interact with the surrounding context. For example, for expressions of type `FORMULA`, the relevant substructures are obtained by splitting the expression at the top node. As a result, we obtain two alternative readings of the expression: $\mathtt{TERM}_1$ `O_FORMULA`$_1$ and `FORMULA_0`$_2$ $\mathtt{TERM}_2$, each of which we again embed within the original text and re-interpret (i.e. re-parse) in the context of the surrounding natural language. The lexical entry for `0_FORMULA` (formula missing an argument to the left) is of syntactic category `s\n` (and its semantics is `such that TERM has property FORMULA`), while the entry for `FORMULA_0` (formula missing an argument to the right) is of category `s/n`. This re-interpretation of a mathematical expression allows us to obtain the intended reading of (4).

Domain-specific syntax With (7), we illustrated the use of domain-specific syntax while verbalising a formal expression in natural language. The past participle

"vereinigt" ("unified") is normally used with a prepositional phrase ("vereinigen mit" + Dat.). The construction "A vereinigt B" presented in this example is, however, commonly used in informal formula verbalisation.[10] To account for this kind of domain-specific constructions, we introduce appropriate syntactic categories for domain-specific lexica in the parser's lexicon. In this case, the lexical entry for "vereinigt" includes a reading as a mathematical operator: `O_TERM_O`, with the syntactic category `s\n/n`. This is similar to the treatment of the operator $\in$, in (2), discussed above.

5.2 Domain Modelling

Imprecision and ambiguity are intrinsic phenomena in natural language. By contrast, formal domains do not tolerate imprecision. In the examples (5) and (7) on p. 271, we presented ambiguous natural language descriptions of the intended SUBSET relation. Another aspect of natural language is its informality; (8) is a verbose description of empty intersection of two sets. Ambiguity, imprecision, and informality require extensions to the basic processing architecture.

We introduce ambiguous concepts into the ontology by representing them as relations that subsume concepts related to mathematical relations. Ambiguous lexical items are linked to ambiguous concepts they evoke through the semantic lexicon. The concepts, in turn, are given an interpretation through the domain ontology.

The semantic lexicon, introduced in Sect. 4, provides a mapping from dependency frames output by the parser to domain-independent conceptual frames. The input structures are described in terms of tectogrammatical valency frames of the lexical items that evoke a given concept. The output structures are either the evoked concepts with roles indexed by tectogrammatical frame elements or results of executing interpretation scripts. Where relevant, sortal information for role fillers is given. Figure 3 shows example lexicon entries explained below. The first two entries are concept evoking; the following are examples of interpretation scripts.

Containment The *containment* relation is evoked, among others, by the verb "enthalten" ("to contain"). The tectogrammatical frame of "enthalten" comprises the relations of `Actor` (*act*) and `Patient` (*pat*) that fill the *container* and the *contents* roles, respectively. The *Containment* relation in its most common domain interpretation specialises into the domain relations of (STRICT) SUBSET or ELEMENT. This specialisation is encoded in the domain ontology. Another kind of containment relation in the sense of *Structural composition* holds between a *structured object* and its structural *sub-component*; the `Patient` dependent of *enthalten* fills the *substructure* role, while the `Actor` dependent is the *Structured object* that embeds the substructure.

Common property The semantics of a general notion of *Common property* is derived using interpretation scripts based on the evoking lexical item "gemeinsam"

[10] Similarly to the English verbalisation of the expression $A \cap B$ as "A union B" instead of the longer, grammatically correct, "the union of A and B".

$$(\textbf{contain}_{pred}, x_{act}, y_{pat}) \rightarrow (\text{CONTAINMENT}_{pred}, \text{container}_{act}, \text{contents}_{pat}) \quad \text{(a)}$$

$$(\textbf{contain}_{pred}, x_{act:formula}, y_{pat:formula}) \rightarrow (\text{STRUCTURAL COMPOSITION}_{pred}, \text{structured object}_{act}, \text{substructure}_{pat}) \quad \text{(b)}$$

$$(\textbf{common}, p_{pred,sem==\textbf{have}}, x_{act:coord(x_1,x_2,\ldots,x_n)}, y_{pat:Pred}) \rightarrow (\text{Pred}(x_1,y_1) \wedge \text{Pred}(x_2, y_1) \wedge \ldots \wedge \text{Pred}(x_n, y_1)) \quad \text{(c)}$$

$$(\textbf{common}, p_{pred,sem:Pred}, x_{act:coord(x_1,x_2,\ldots,x_n)}, y_{pat}) \rightarrow (\text{Pred}(x_1, y) \wedge \text{Pred}(x_2, y) \wedge \ldots \wedge \text{Pred}(x_n, y)) \quad \text{(d)}$$

$$(\textbf{common}, p_{pred,sem:Pred1}, x_{act:coord(x_1,x_2,\ldots,x_n)}, y_{pat:Pred2}) \rightarrow (\text{Pred1}(x_1, y_1) \wedge \text{Pred1}(x_2, y_1) \wedge \ldots \wedge \text{Pred1}(x_n, y_1) \wedge \text{Pred2}(x_1, y_1) \wedge \text{Pred2}(x_2, y_1) \wedge \ldots \wedge \text{Pred2}(x_n, y_1)) \quad \text{(e)}$$

Fig. 3 Example entries from the semantic lexicon

("common"). Three such scripts are presented in Fig. 3. *Pred* is an object which can be instantiated with a relational noun, such as "an element (of)", in the first entry, (c). Using the interpretation script, the representation of the utterance "*A* and *B* have no common elements" can be instantiated as follows: **common**, $p_{pred,sem==\textbf{have}}$, $x_{act:coord(A,B)}$,$y_{pat:Element} \rightarrow$ (ELEMENT(A,y_1) ELEMENT(B, y_1). Entry (d) serves to interpret utterances containing a relational predicate whose `Patient` dependent is a common noun, such as "[Peter and Paul]$_{act:coord(x_1,\ x_2)}$ [have]$_{pred,\ sem:Pred}$ [a common car]$_{pat}$". The last entry, (e), is the most general case for utterances containing both a relational noun and a relational predicate, such as "[Peter and Paul]$_{act:coord(x_1,\ x_2)}$ [see]$_{pred,\ sem:Pred1}$ [a common friend]$_{pat:Pred2}$".

The domain ontology is an intermediate representation that provides a link to logical definitions in a mathematical knowledge base and mediates between the discrepant views of linguistic analysis and deduction systems' representation (cf. [22]). Its most interesting feature is the representation of ambiguous and general concepts, such as *Containment* (cf. above), as *semantic relations* in the ontology of relations. In terms of the associated semantics, we view them as subsuming mathematical relations. For example, a *Containment* holds between two entities if one includes the other as a whole, or all its components separately. This is a generalisation of the (STRICT) SUBSET and ELEMENT relations in set theory.

5.3 Domain-Specific Anaphora

Anaphoric devices are used to refer to the formal expressions or to their parts. Aside from the most obvious anaphoric expressions, such as "the formula" or "the term", other naturally occurring references include "the left/right side" (of a term or formula), and "the (left/right) bracket". To account for discourse references to parts of mathematical expressions, two issues have to be taken into account: first, the relevant substructures of mathematical expressions must be identified and second, they must be included in the domain ontology. We identified the relevant mathematical expression parts empirically by corpus analysis and observations on usage. Notably, we account for references to types of expressions (e.g. "the formula", "the term"),

typographical features, i.e. physical properties of the expressions (such as "sides" of terms and formula), linear orders (e.g. "first", "second" argument), and structural groupings (delimited sub-expressions, such as "bracket", "inner bracket").

We extended the domain ontology of objects to include information on mathematical expression substructures (e.g. a "side" of a formula is a mathematical expression itself, and it is of type `TERM`). Procedural functionality needed to resolve references such as "left side" or "inner parenthesis" involved extending the mathematical expression parser by functions that recover specific substructures of mathematical expressions in specific `part-of` relations to the original expression.

As mentioned previously, some of the references have a metonymic flavour. For example, by saying "the left side" of a formula, we do not mean the side as such, but rather the term to the given side of the main operator in the expression. The use of such metonymic expressions is so common and systematic in the mathematical terminology that it is justified to consider them as quasi-synonyms of the expressions they refer to. Following this systematicity, we encoded polysemy rules for treatment of metonymic references, as part of the domain model. Examples of such rules in the mathematics domain are `formula-side : term-at-side` (meaning that a "side" of a formula can be interpreted as the term to the given side of the top operator), as in (11); `bracket : bracketed-term` (a "bracket" is to be interpreted as a term enclosed by a pair of brackets); and `operator : term-under-operator` (an operator is to be interpreted as a term headed by the given operator), as in (13).

5.4 Flexible Formula Analysis and Disambiguation

In formal domains, students are prone to low-level errors while attempting to construct formal expressions. In principle, in a dialogue environment, clarification subdialogues could be initiated to indicate imprecision or error, and to elicit clarification or correction, respectively. This may, however, turn out unwieldy and tedious and therefore particularly undesirable when the problem solving skills of the student are otherwise satisfactory. A better solution is to attempt to cooperatively correct what appears to be an error or to resolve ambiguity, while allowing the student to concentrate on the problem solving itself.

Using domain knowledge, and in particular reasoning about the proof state, erroneous or ambiguous mathematical statements may be evaluated for correctness and relevance in a given proof state. However, identification of the intended interpretation and the decision as to whether the flawed statement should be corrected by the student is pragmatically influenced by other factors: for example, information on the student's knowledge of the domain concepts and their correct use, correct usage of the domain terminology or contextual preference for one reading over the others. On the one hand, in the most "accommodating" approach, erroneous and ambiguous expressions evaluated as correct in one of the readings could be accepted without requiring clarification on the part of the student, thus making the dialogue

progression smoother. On the other hand, in a tutoring context, it is of utmost importance to identify the student's intention correctly.

Our goal is to delay clarification, while making sure that the student's intentions remain tractable. To decide whether to accept an erroneous or ambiguous utterance (a strategy suitable for competent students) or to issue a clarification request for the student to disambiguate the utterance explicitly, we can use the adopted teaching strategy and the student model maintained by the Tutorial Manager [37].

In case of flawed mathematical expressions, such as those exemplified by (14)–(17) and (19) on p. 273, we attempt to identify and cooperatively correct the error. To this end, we built a flexible mathematical expression analyser and a formula correction module. When we encounter an ill-formed expression, we attempt to recover the correct (possibly intended) expression by applying local and contextually justified modifications. The flexible mathematical expression analysis enables cooperative reactions in the sense that the statement which was probably intended is hypothetically assumed and interpreted in the problem-solving context, so that its correctness and relevance can be addressed, while the fact that it was ill-formed can be merely signalled to the student, pointing at the error.

Finding meaningful modifications to an expression that aim at reconstructing its corrected and possibly intended variant is guided by the expression's *correctness state*. We distinguish three categories of errors: (1) *logical* errors (e.g. too weak or too strong statements in the given context), (2) *type* errors (types of operands do not match the types required by the given operator), and (3) *structural* errors (syntactic errors). Categorisation of well-formed mathematical expressions as to their correctness status is obtained from the Proof Manager. However, the formula analyser may not be able to analyse the given expression on the basis of the provided constructors.

Flexible formula analysis relies on associating a set of replacement rules with each error category and applying the rules trying to achieve an improvement at least one category level as a result of the modification; e.g. from an ill-formed expression we obtain a well-formed expression, and from an expression with a type mismatch we obtain a well-typed expression. Examples of replacements rules include and dual operators, e.g. $\cap \Leftrightarrow \cup$, $\subset \Leftrightarrow \supset$; stronger/weaker operators, e.g. $\supset \Leftrightarrow \supseteq$, $\subseteq \Leftrightarrow =$; confusable operators or identifiers, eg $\subset \Leftrightarrow \in$, $K \Leftrightarrow P$, $p \Leftrightarrow b$; and insert blank or brackets, e.g. $PC \Rightarrow P\ C$, $PC \Rightarrow P(C)$. To constrain the application of changes, we assume a context consisting of a set of identifiers (i.e., variables and operators), together with type and operator arity information. Each rule is justified by the evidence of the plausibility of the given error occurring in the domain, the nature of the task, and student's capabilities. More details on the flexible formula analysis as described above can be found in [20].

6 Modelling Dialogue for Mathematics Tutoring

The flexible analysis of students' statements outlined in the previous section is an example of how the modules within our tutorial dialogue system cooperate to analyse inputs and to decide what responses should be given. The analysis takes

place at multiple levels, and different inputs require the modules to interact in different ways. For instance, only when the natural language understanding module determines that an utterance was a domain contribution is it sent to the Proof Manager to evaluate its domain content. Facilitating this multi-layered analysis is a requirement for the design of the dialogue management process. The design of dialogue models must also take into account certain characteristic features of tutorial dialogue. For instance, tutors and students do not have the same obligations arising from questions, and tutors may choose not to answer questions whereas students must. Also, repeated information has different effects on the mutual beliefs of the dialogue participants [24].

Both of these requirements – module cooperation and the characteristics of tutorial dialogue – influence our design choices when building a conversational agent for tutoring. It maintains a model of the dialogue state, and, given an analysis of an incoming utterance along with conversational expertise, determines an appropriate response. The dialogue model is continually updated to reflect the effect of both system and user utterances, and in turn forms the context for future response choices. The dialogue model representation is influenced by the topic of the dialogue, i.e. what objects are being talked about. Similarly, how the model is updated depends on the pragmatic conditions of the dialogue, in other words, how students and tutors should behave.

Our initial work focused on models for dialogue management. A dialogue manager, as introduced in Sect. 2, contains the dialogue model and facilitates communication between other modules. Our prototype tool [9] used an agent-based blackboard architecture to allow system modules to share the information stored in the dialogue model. Updates were carried out by software agents which monitored the state of the dialogue model. However, the strengths of this prototype went beyond the needs of a dialogue management application, and did not justify its complexity compared to other tools. Our continuing work on dialogue modelling therefore uses Dipper [8], a standard rule-based dialogue management tool.

Integrated modelling of dialogue state In a tutorial dialogue system, the choice of system action depends on the analysis results from a number of system modules. This motivates our use of a combined model of the state of the dialogue consisting of three levels: linguistic/pragmatic information, proof state information, and tutoring state information. In [10] we have proposed a model of how these three information sources can be combined into a single representation of dialogue state. The proof level information abstracts from the task model stored in the Proof Manager. It includes the exercise being tutored, its current status, a proof step history, and a representation of the last proof step. The information which describes the last proof step is the result of the analysis of the utterance by both the NLU module and the Proof Manager, and includes the formula which was uttered and the type of the proof step as well as the proof step evaluation.

The model of tutorial state maintains features identified in [37], for instance the number and type of hints that have been given, and the number and type of errors that the student has made. The linguistic level information is captured in a simple

```
rule( evaluateDomCon,
 [ $/dmodel/lu/speaker = student,
   in($/dmodel/lu/moves, domcon),
   $/taskcontext/lps/steptype = Type,
   $/taskcontext/lps/content = Formulas,
   proofmanager_interface(evaluate,[Type,Formulas,Evaluation])
 ],
 [ set(/taskcontext/lps/evaluation,Evaluation)]).
```

Fig. 4 An update rule combining dialogue and task level information

dialogue model which includes a dialogue history, a representation of the last utterance as delivered by the NLU module and outlined in Sect. 4, and a stack of moves which the system intends to express in its next turn.

Updates to the dialogue state are encoded in update rules consisting of preconditions and effects. An example is shown in Fig. 4, which states that if the student's utterance is a domain contribution, its domain content should be evaluated by the Proof Manager and the evaluation of the step should be added to the task model. This rule shows how dialogue level and task level information combine to trigger an update to the model of dialogue state as a whole.

Modelling Common Ground One of the characteristics of tutorial dialogue which is important for building ITSs is the effect of tutors' evaluations of students' contributions. The tutor's authority to "decree" the truth of assertions should cause students to believe that those assertions which the tutor has accepted are indeed true. In the dialogue model we capture this by modelling the *common ground* [13], the set of propositions which dialogue participants are said to mutually believe as a result of their dialogue. Information becomes part of the common ground by being presented by a speaker and subsequently accepted by a hearer, a process known as *grounding* [13, 35]. Problems in the grounding process can lead to dialogue participants having differing beliefs about the common ground. This situation of *misalignment* can lead to subsequent interpretation problems.

In [11] we have presented a model of common ground for tutorial dialogue in which the common ground develops as a store of the truth and falsity of the domain contributions that the student has made. The model allows us to detect *informationally redundant utterances* (IRU) [41] and use this information (1) to identify possible misalignment and (2) to inform the tutorial manager that the misalignment occurred, so that it can decide to apply pedagogical strategies to remedy the misalignment. In example (20) the student shows the truth of the formula $(b, a) \in (R \circ S)$ in utterance S3, which is confirmed by the tutor and thereby becomes grounded. In utterance S6 the student assumes the same formula; S6 is thus an IRU. Since S6 is not marked to show that the repetition is intentional, the tutor can conclude that the student does not consider the truth of the formula to be part of their common ground. This is evidence of misalignment, which causes the tutor to produce the hint given in T6 with the goal of realigning the student's beliefs.

(20) *S3:* Let $(a, b) \in (R \circ S)^{-1}$. Then it holds that $(b, a) \in (R \circ S)$
T3: That's right.
...
S6: Let $(b, a) \in (R \circ S)$. Then ...
T6: Since you already know that $(b, a) \in (R \circ S)$, you don't need to postulate it again.

We also use the model of common ground to inform the natural language generation module to mark utterances which contain information which is to be intentionally repeated. It is often necessary to repeat previously grounded information, especially in task-oriented dialogue, for instance as a reminder or to make a referent salient again. Such utterances, which are IRUs, must be marked as such so that the hearer does not falsely conclude that misalignment has taken place. In example (21), after $A \cap B = \emptyset$ has been grounded earlier in the dialogue, the particle "of course" marks the fact that (part of) T8 is being intentionally repeated.

(21) *S2:* $A \cap B = \emptyset$
...
T8: ... The justification could for instance be: ... (since of course $A \cap B = \emptyset$) ...

By comparing the current content of the common ground with the content of an utterance which is to be generated by the system, the model allows us to decide whether marking for informational redundancy should be added. This supports the process of cognitive alignment between student and tutor.

In summary, by monitoring the common ground of the interaction the system can detect when misalignment has occurred in cases where it is explicitly linguistically marked and can also mark its own utterances in order to avoid the student falsely concluding that misalignment has occurred.

7 Related Work

Language understanding in practical dialogue systems, be it with text or speech interface, is commonly performed using shallow syntactic analysis combined with keyword spotting. Intelligent Tutoring Systems also successfully employ statistical methods which compare student responses to a model built from pre-constructed gold-standard answers [17]. When precise understanding is needed, some tutorial systems either use menu- or template-based input (e.g. [18]), or use closed-questions to elicit short answers of little syntactic variation [15]. These approaches are insufficient in our scenario because, on the one hand, we need in-depth understanding of the students' input, and on the other hand, we want to give them freedom in expressing their ideas.

Several recent Intelligent Tutoring Systems addressing subfields of formal domains, for instance elementary geometry (PACT [32]), electrical engineering (BEETLE [48]) and qualitative physics (Why2-Atlas [25]), offer natural

language-based interaction. Interpretation capabilities in those systems are considerably robust (cf. [33, 14]). However, unlike in our case, the input to those systems is characterised by only a limited range of mixed language phenomena (see [23, 26]).

Baur [3] and Zinn [46] present DRT-based analyses of selected linguistic phenomena in example course-book proofs. However, the language in dialogues is more informal than in course-books. Both above approaches rely on typesetting and additional information that identifies mathematical symbols, formulas, and proof steps. Forcing the user to delimit formulas would, however, reduce the flexibility of the system and make the interface harder to use, while not guaranteeing a clean separation of the natural language and the non-linguistic content anyway (as is the case of the LeActiveMath system whose interface attempts to separate the symbolic content from the linguistic content [12]). The work presented here addresses both aspects mentioned above: an informal mathematical discourse that is additionally placed in a dialogue setting.

Dialogue based interaction in tutoring has been shown to be more effective than less interactive instruction [29, 39]. Moreover, grounding has been identified as an important factor in both peer learning [2] and tutoring [30]. Intelligent tutoring systems, such as AutoTutor [31] and Ms Lindquist [19], which use simple dialogue models, have no model of common ground, but capture misconceptions using explicit rules. In those systems, there is no clear separation between modelling the dialogue itself and modelling the tutoring task. The dialogue advances according to the local tutoring agenda. Zinn [47] presents a dialogue-based tutoring system in which discourse obligations are generated from a store of task solution descriptions and the common ground is maintained in the dialogue model. However, the choice of tutoring actions is not informed by the state of the common ground, but rather is explicitly encoded. Our work addresses the shortcomings of these approaches by using the dialogue model to inform the choice of pedagogical actions while keeping dialogue expertise separate from pedagogical knowledge.

8 Conclusions

We presented interpretation methods for major phenomena found in mathematical natural language utterances in a corpus of simulated human–computer tutorial dialogues on theorem proving. Discourse contributions found in this corpus combine natural language descriptions and mathematical expressions in an interleaved way. Both modes are flawed with incompleteness and errors, which makes them largely inaccessible to present analysis techniques.

Our interpretation approach addresses the mathematical discourse phenomena in a stratified way, starting with basic methods, enhancing them by techniques addressing mixed language, and adding features for handling faulty constructions in a cooperative manner. Interpretation is a step-wise procedure supported by a number of knowledge sources, including not only lexical and grammatical resources, but also domain representations and reasoning tools, in a highly modular architecture.

The central component which enables the interaction between student and tutor is the dialogue manager. We have developed a model of tutorial dialogue which supports the complete dialogue management process and addresses specific phenomena found in this genre. The model takes the analysis of the utterance as input, maintains the dialogue state by integrating information supplied by external modules, and using conversational expertise chooses the system response.

We have implemented prototype input interpretation and dialogue manager modules capable of analysing the phenomena discussed in this chapter. Sub-modules of the interpretation component and the knowledge resources it employs have been implemented and tested on sentences from the corpus and constructed examples. We are presently extending the functionality to account for more complex phenomena.

References

1. Autexier, S., Benzmüller, C., Fiedler, A., Horacek, H., Vo, B.Q. Assertion-level proof representation with under-specification. In: Proceedings of the Mathematical Knowledge Management Symposium, Electronic Notes in Theoretical Computer Science (vol. 93, pp. 5–23) (2004).
2. Baker, M., Hansen, T., Joiner, R., Traum, D. The role of grounding in collaborative learning tasks. In: Dillenbourg, P. (Ed.), Collaborative Learning. Cognitive and Computational Approaches, Advances in Learning and Instruction Series (pp. 31–63). Amsterdam, The Netherlands: Pergamon (1999).
3. Baur, J. Syntax und Semantik Mathematischer Texte. Diplomarbeit, Fachrichtung Computerlinguistik. Saarbrücken, Germany. Universität des Saarlandes (1999).
4. Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Tsovaltzi, D., Vo, B.Q., Wolska, M. A Wizard-of-Oz experiment for tutorial dialogues in mathematics. In V. Aleven, U. Hoppe, J. Kay, R. Mizoguchi, H. Pain, F. Verdejo, K. Yacef, (Eds.), AIED-03 Supplementary Proceedings, Advanced Technologies for Mathematics Education (vol. III, pp. 471–481). Australia: Sydney (2003).
5. Benzmüller, C., Horacek, H., Kruijff-Korbayová, I., Lesourd, H., Schiller, M., Wolska, M. DiaWozII – A tool for Wizard-of-Oz experiments in mathematics. In: Proceedings of the 29th Annual German Conference on Artificial Intelligence (KI-06), Lecture Notes in Computer Science (vol. 4314, pp. 159–173). Springer-Verlag, Bremen, Germany (2006).
6. Benzmüller, C., Horacek, H., Lesourd, H., Kruijff-Korbayová, I., Schiller, M., Wolska, M. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06) (pp. 1766–1769). ELDA, Genoa, Italy (2006).
7. Benzmüller, C., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Wolska, M. Natural language dialog with a tutor system for mathematical proofs. In R. Lu, J. Siekmann, C. Ullrich, (Eds.), Cognitive Systems, Lecture Notes in Computer Science (vol. 4429, pp. 1–14). New York: Springer (2007).
8. Bos, J., Klein, E., Lemon, O., Oka, T. Dipper: Description and formalisation of an information-state update dialogue system architecture. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue (pp. 115–124) (2003).
9. Buckley, M., Benzmüller, C. An Agent-based Architecture for Dialogue Systems. In I. Virbitskaite, A. Voronkov, (Eds.), Proceedings of Perspectives of System Informatics, Lecture Notes in Computer Science (vol. 4378, pp. 135–147). Novosibirsk, Russia: Springer (2006).
10. Buckley, M., Dietrich, D. Integrating task information into the dialogue context for natural language mathematics tutoring. In B. Medlock, D. Ó Séaghdha, (Eds.), Proceedings of the 10th Annual CLUK Research Colloquium. Cambridge, UK: University of Cambridge (2007).

11. Buckley, M., Wolska, M. Towards modelling and using common ground in tutorial dialogue. In R. Artstein, L. Vieu, (Eds.), Proceedings of DECALOG, the 2007 Workshop on the Semantics and Pragmatics of Dialogue (pp. 41–48). Rovereto, Italy (2007).
12. Callaway, C., Dzikovska, M., Matheson, C., Moore, J., Zinn, C. Using Dialogue to Learn Math in the LeActiveMath Project. In: Proceedings of the ECAI-06 Workshop on Language-enabled Educational Technology (pp. 1–8). Riva del Garda, Italy (2006).
13. Clark, H.H., Schaefer, E.F. Contributing to discourse. Cognitive Science, 13:259–294 (1989).
14. Core, M.G., Moore, J.D. Robustness versus fidelity in natural language understanding. In R. Porzel, (Ed.), HLT-NAACL-04 Workshop: 2nd Workshop on Scalable Natural Language Understanding (pp. 1–8). Boston, Massachusetts, USA: Association for Computational Linguistics (2004).
15. Glass, M. Processing language input in the CIRCSIM-tutor intelligent tutoring system. In: Proceedings of the 10th International Conference on Artificial Intelligence in Education (AIED-01) (pp. 210–221). IOS Press, San Antonio (2001).
16. Graesser, A.C., Person, N.K., Magliano, J.P. Collaborative dialogue patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology, 9:495–522 (1995).
17. Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8:129–147 (2000).
18. Heffernan, N.T., Koedinger, K.R. Intelligent tutoring systems are missing the tutor: Building a more strategic dialog-based tutor. In C. Rosé, R. Freedman, (Eds.), Proceedings of the AAAI Fall Symposium on Building Dialogue Systems for Tutorial Applications (pp. 14–19). Menlo Park, CA: AAAI Press (2000).
19. Heffernan, N.T., Koedinger, K.R. An intelligent tutoring system incorporating a model of an experienced human tutor. In: Proceedings of the 6th International Conference on Intelligent Tutoring Systems (pp. 596–608). London, UK (2002).
20. Horacek, H., Wolska, M. Fault-tolerant context-based interpretation of mathematical formulas. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05) (pp. 1688–1689). Lawrence Erlbaum Associates Ltd, Edinburgh, Scotland (2005).
21. Horacek, H., Wolska, M. Interpreting semi-formal utterances in dialogs about mathematical proofs. Data & Knowledge Engineering, 58(1):90–106 (2006).
22. Horacek, H., Fiedler, A., Franke, A., Moschner, M., Pollet, M., Sorge, V. Representation of mathematical objects for inferencing and for presentation purposes. In: Proceedings of EMCSR-2004 (pp. 683–688). Vienna, Austria (2004).
23. Jordan, P.W., Makatchev, M., Pappuswamy, U. Relating student text to ideal proofs: Issues of efficiency of expression. In: Proceedings of the AIED-05 Workshop on Mixed Language Explanations in Learning Environments, Amsterdam (pp. 43–50). The Netherlands (2005).
24. Karagjosova, E. Marked informationally redundant utterances in tutorial dialogue. In: Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck) Saarbrücken (2003).
25. Makatchev, M., Jordan, P., VanLehn, K. Modeling students' reasoning about qualitative Physics: Heuristics for Abductive Proof Search. In: Proceedings of the International Conference on intelligent tutorial systems (ITS 2004), Lecture Notes in Computer Science (vol. 3220, pp. 699–709). Springer (2004).
26. Makatchev, M., Hall, B.S., Jordan, P.W., Pappuswamy, U., VanLehen, K. Mixed language processing in the why2-atlas tutoring system. In: Proceedings of the AIED-05 Workshop on Mixed Language Explanations in Learning Environments (pp. 35–42). Amsterdam, The Netherlands (2005).
27. McArthur, D., Stasz, C., Zmuidzinas, M. Tutoring techniques in algebra. Cognition and Instruction, 7(3):197–244 (1990).
28. Merrill, D.C., Reiser, B.J., Merrill, S.K., Landes, S. Tutoring: Guided learning by doing. Cognition and Instruction, 13:315–372 (1995).

29. Moore, J. What makes human explanations effective? In: Proceeding of the 15th Annual Conference of the Cognitive Science Society (pp. 131–136). Hillsdale, NJ (1993).
30. Pata, K., Sarapuu, T., Archee, R. Collaborative scaffolding in synchronous environment: Congruity and antagonism of tutor/student facilitation acts. In T. Koschman, D. Suthers, T.W. Chan, (Eds.), Computer Supported Collaborative Learning 2005: The next 10 years (pp. 484–493). Dordrecht: Kluwer (2005).
31. Person, N.K., Bautista, L., Kreuz, R.J., Graesser, A.C. The Tutoring Research Group. The dialog advancer network: A conversation manager for autotutor. In: Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies at the 5th International Conference on Intelligent Tutoring Systems (ITS-00) (pp. 86–92). Montreal, Canada (2000).
32. Popescu, O., Koedinger, K.R. Towards understanding geometry explanations. In: Building Dialogue Systems for Tutorial Applications, Papers from the 2000 AAAI Fall Symposium (pp. 80–86). AAAI Press, Menlo Park, California (2000).
33. Rosé, C.P., Gaidos, A., Hall, B.S., Roque, A., VanLehn, K. Overcoming the knowledge engineering bottleneck for understanding student language input. In: Proceedings of the 11th International Conference on Artificial Intelligence in Education (AIED-03) (pp. 315–322). Sydney, Australia (2003).
34. Sgall, P., Hajičová, E., Panevová, J. The Meaning of the sentence in its Semantic and Pragmatic Aspects. Dordrecht, The Netherlands: Reidel Publishing Company (1986).
35. Traum, D. Computational models of grounding in collaborative systems. In S.E. Brennan, A. Giboin, D. Traum, (Eds.), Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, AAAI (pp. 124–131). California: Menlo Park (1999).
36. Traum, D., Larsson, S. The information state approach to dialogue management. In J. van Kuppevelt, R. Smith, (Eds.), Current and New Directions in Discourse and Dialogue (pp. 325–354). Berlin, Germany: Kluwer (2003).
37. Tsovaltzi, D., Fiedler, A. Human-adaptive determination of natural language hints. In: Proceedings of CMNA 5, 5th Workshop on Computational Models of Natural Argument, IJCAI-05 (pp. 84–88) (2005).
38. Tsovaltzi, D., Horacek, H., Fiedler, A. Building hint specifications in a NL tutorial system for mathematics. In: Proceedings of the 16th International Florida AI Research Society Conference (FLAIRS-04) (pp. 929–934). Florida, USA (2004).
39. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rosé, C.P. When are tutorial dialogues more effective than reading? Cognitive Science: A Multidisciplinary Journal 31(1):3–62 (2007).
40. Vo, B.Q., Benzmüller, C., Autexier, S. Assertion application in theorem proving and proof planning. In: Proceedings of the 10th Workshop on Automated Reasoning: Bridging the Gap between Theory and Practice. Liverpool, England (2003).
41. Walker, M. Informational redundancy and resource bounds in dialogue. PhD thesis, University of Pennsylvania, Philadelphia, PA (1993).
42. Wolska, M., Kruijff-Korbayová, I. Analysis of mixed natural and symbolic language input in mathematical dialogs. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL-04) (pp. 25–32). Barcelona, Spain (2004).
43. Wolska, M., Kruijff-Korbayová, I. Factors influencing input styles in tutoring systems: The case of the study-material presentation format. In: Proceedings of the ECAI-06 Workshop on Language-enabled Educational Technology (pp. 86–91). Riva del Garda, Italy (2006).
44. Wolska, M., Kruijff-Korbayová, I. Modeling anaphora in informal mathematical dialogue. In: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (brandial-06) (pp. 147–154). Potsdam, Germany (2006).
45. Wolska, M., Vo, B.Q., Tsovaltzi, D., Kruijff-Korbayová, I., Karagjosova, E., Horacek, H., Gabsdil, M., Fiedler, A., Benzmüller, C. An annotated corpus of tutorial dialogs on mathematical theorem proving. In: Proceedings of International Conference on Language Resources and Evaluation (LREC-04). ELDA, Lisbon, Potugal (2004).

46. Zinn, C. A computational framework for understanding mathematical discourse. Logic Journal of the IGPL, 11(4):457–484 (2003).
47. Zinn, C. Flexible dialogue management in natural-language enhanced tutoring. In: Konvens 2004 Workshop on Advanced Topics in Modeling Natural Language Dialog (pp. 28–35). Vienna, Austria (2004).
48. Zinn, C., Moore, J.D., Core, M.G., Varges, S., Porayska-Pomsta, K. The BE&E tutorial learning environment (BEETLE). In: Proceedings of the Seventh Workshop on the Semantics and Pragmatics of Dialogue (DiaBruck). Saarbrucken, Germany (2003).

Resource-Bounded Modelling and Analysis of Human-Level Interactive Proofs

Christoph Benzmüller, Marvin Schiller, and Jörg Siekmann

1 Introduction

Mathematics is the *lingua franca* of modern science, not least because of its conciseness and abstractive power. The ability to prove mathematical theorems is a key prerequisite in many fields of modern science, and the training of how to do proofs therefore plays a major part in the education of students in these subjects. Computer-supported learning is an increasingly important form of study since it allows for independent learning and individualised instruction.

Our research aims at partially automating intelligent tutoring of mathematical proofs. This research direction is interesting not least because of the large number of potential users of such systems, including students who in addition to an introductory university lecture want to exercise their theorem proving skills, learners without access to university courses, and engineers who want to freshen their skills. Furthermore, the research direction is interesting because of the non-trivial challenge it poses to artificial intelligence, computational linguistics and e-learning: in order to achieve a powerful and effective intelligent proof tutoring system many research problems that are central to these areas have to be addressed and combined.

In the SFB 378 project DIALOG [12, 6, 8] (see also Wolska *et al.*, Linguistic Processing in a Mathematics Tutoring System of this volume) we have revealed and addressed foundational research challenges that are crucial for realising intelligent computer-supported proof tutoring based on a flexible, natural language-based dialogue between student and computer. In the proof tutoring scenario as studied in the project the student communicates proof steps to the tutorial system by embedding them in natural language utterances. The language used is a mixture of natural-language and mathematical expressions ("mathural" [27]). Proof construction is performed in a stepwise fashion, and the system responds to utterances with appropriate didactically useful feedback or also with hints. The student is free to build any valid proof of the theorem at hand.

C. Benzmüller (✉)
Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany
e-mail: chris@ags.uni-sb.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_13,

To support the generation of appropriate feedback each proposed proof step needs to be analysed by the system in the context of the partial proof developed so far. For this reason, automating proof tutoring requires dynamic techniques that assess the student's proof steps on a case-by-case basis in order to generate the appropriate feedback. The feedback can take the form of confirming correct steps, drawing the student's attention to errors and offering domain specific hints when the student gets stuck. In case the tutor system is asked to give a hint, the hint is generated in the context of the current proof, and it has to be exactly tailored to the situation in which the hint was requested. The ability to dynamically construct proofs, to dynamically analyse new proof steps, and to complete partial proofs to full proofs is thus an essential prerequisite for intelligent proof tutoring.

The scenario we finally envisage integrates the flexible, dialogue-based proof tutoring system we are aiming at with an interactive e-learning environment for mathematics. An example of an interactive e-learning environment is ActiveMath [16]. ActiveMath is a third-generation e-learning system for school and university level learning as well as for self-study that offers new ways to learn mathematics. In ActiveMath the learner can, for example, choose among several learning scenarios, receive learning material tailored to his/her needs and interests, assemble individual courses himself/herself, learn interactively and receive feedback in exercises, use interactive tools, and inspect the learner model and partially modify the system's beliefs about the student's capabilities and preferences. The flexible, dialogue-based proof tutoring system which we aim at shall ideally cooperate with such an e-learning environment. A learner taking an interactive course in the e-learning system shall be able to call it in order to exercise his/her theorem proving skills within the trained mathematical domain. Ideally, both the e-learning environment and the proof tutoring system share the formal mathematical content, the didactic goals and the student model. The exercise within the proof tutoring environment will then exploit this information and confirm, modify or refine the student model.

The combination of expertise from computational linguistics and from deduction systems made the research in the DIALOG project particularly interesting. Expertise from the former area was needed because of the choice of the flexible mathural language as communication means between student and system. Expertise in the latter area was needed for the development of techniques for dynamic proof management and dynamic proof step evaluation.

The remainder of the paper is organised as follows: In Sect. 2, we illustrate the initial position of our project, where the relative lack of data prompted empirical investigations. Based on the collected data, we formulate research challenges for proof tutoring in Sect. 3. A central role among those challenges is dynamic proof step evaluation, which is approached in Sect. 4. Mathural processing is the subject of the article by Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring System of this volume) and human-oriented theorem proving in Ω*mega* is explained in the article by Autexier *et al.* (Resource-Adaptive Processes in Automated Reasoning Systems in this volume). This chapter can be seen as a bridge between these two articles. Section 5 elaborates on didactic strategies, dialogue modelling, feedback generation and hints. Section 6 concludes the article and relates our work to other approaches in the field.

2 The Need for Experiments and Corpora

In order to make a start in this research direction, experiments were needed to obtain corpora that could guide our foundational research. Not only was little known about the type of natural language utterances used in student – tutor dialogues on proofs but also there was little work available about automating proof tutoring based on flexible student – tutor dialogues. To collect a corpus of data, which now forms the basis of our investigations into dialogue-based proof tutoring, we conducted two experiments in the Wizard-of-Oz style, which included the work of Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring System of this volume).

Experiment 1: A first experiment [7] served to collect a corpus of tutorial dialogues in the domain of proof tutoring. It investigated the correspondence between domain-specific content and its linguistic realisation, and the use, distribution and linguistic realisation of dialogue moves in the mathematics domain. It also investigated three tutoring strategies, a *Socratic* tutoring strategy (cf. [32]), a *didactic* strategy and a *minimal feedback* strategy (where the subjects only obtained very brief feedback on the correctness of their attempts).

Setup. A tutorial dialogue system was simulated in the Wizard-of-Oz paradigm [24], i.e. with the help of a human expert. Twenty-four university students were instructed to evaluate the dialogue system. Their task was to solve exercises from naive set theory in collaboration with the system. The communication between student and tutor, who was hidden in a separate room, was mediated with a software tool DiaWoZ [18], which was specifically designed for that purpose. A comfortable and usable interface is important for three reasons: (i) in a Wizard-of-Oz setting, the tutor is more efficient in constructing responses and thus better able to conceal his identity, (ii) the system as such appears more mature (and thus plausible) to the student, which helps to further disguise the Wizard-of-Oz setup and (iii) in general, it allows a more fluent interaction with the system (and thus the student to concentrate on the learning activity). We like to thank the department of computer linguistics who provided the tools and facilities for this kind of experiments.

DiaWoZ supports complex dialogue specifications, which were needed in order to specify a particular hinting algorithm used in the *Socratic* tutoring condition. DiaWoZ allows keyboard-to-keyboard interaction between the wizard and the student. The interfaces consist mainly of a text window with the dialogue history and a menu bar providing mathematical symbols. Furthermore, the wizard can assign dialogue state transitions and speech act categories to student turns w.r.t. the underlying dialogue model. The DiaWoZ interface allowed a free mix of natural language text with mathematical symbols. However, we still found a number of aspects of the system which we were able to improve. For example, the participants suggested the use of the keyboard instead of the mouse for inserting mathematical symbols. Therefore, we have gradually improved our experiment software w.r.t. its interface components.

Corpus. The corpus of dialogues between students and the tutor contains data of 22 subject, i.e. 22 experiment sessions. Each session is subdivided into three

subdialogues w.r.t. to three different exercises that the students were required to work on in sequence. These 66 subdialogues contain on average 12 dialogue turns. In total, the corpus consists of 1,115 dialogue turns (of which 393 are student turns), where an average student turn consists of one sentence, and an average tutor turn consists of two sentences. The example dialogue excerpt in Fig. 1 shows tutoring-related annotations (discussed further in [42] and [6]). The tutor turns are annotated with hint categories, the student turns are annotated with the student answer categories.

T1 Please show: If $A \subseteq K(B)$, then $B \subseteq K(A)$!
S1 (wrong) $A \subseteq B$
T2: (give-away-relevant-concept) That is not correct! First you have to consider the if–then – relation.
S2: (wrong) $A \subseteq K(K(A))$
T3: (elaborate-domain-object) That is correct, but at the moment not interesting. Do you know how to deal with the if–then–relation?

Fig. 1 Dialogue excerpt from the first experiment. (K denotes a set complement.)

Experiment 2: A second experiment investigated mathematical domain reasoning tasks and linguistic phenomena in tutorial dialogues. In contrast to the first experiment, it imposed less constraints on the wizards' tutoring and assumes a rather simple dialogue model.

Setup. Thirty-seven students from Saarland University interacted with the mock-up dialogue system, simulated with our software environment DiaWOz-II and four experts,[1] who took the role of the wizard in turn. As a minimal requirement, students were required to have completed at least one university-level mathematics course. The students were instructed to solve mathematical exercises in collaboration with the system. The exercises were taken from the domain of relations and were centred around the concepts of relation composition and relation inverse. Because of the advanced character of the exercises, the participants had to fulfil the prerequisite of having taken part in at least one mathematics course at university level prior to the experiment. At first, the subjects were required to fill out a questionnaire, collecting data about previous experiences with dialogue systems and their mathematics background. Subjects were also given study material with the mathematical definitions that were required to solve the exercises. The largest part of the 2-hour experimental session was allotted to the interaction between the student and the simulated system.

Our WOZ environment DiaWOz-II [10] enables dialogues where natural language text is interleaved with mathematical notation, as is typical for (informal) mathematical proofs. The interface components of DiaWOz-II are based on the *what-you-see-is-what-you-get* scientific text editor TeXmacs [2] [21]. DiaWOz-II

[1] The experts consisted of the lecturer of a course *Foundations of Mathematics*, a maths teacher, and two maths graduates with teaching experience.
[2] www.texmacs.org

provides one interaction window for the user and one for the wizard, together with additional windows displaying instructions and domain material for the user, and additional notes and pre-formulated text fragments for the wizard. All of these windows allow for copying freely from one to the other. Furthermore, our DiaWOz-II allows the wizard to annotate user dialogue turns with their categorisation. DiaWOz-II is also connected to a spell-checker for checking both the user's and the wizard's utterances.

Corpus. The collected corpus contains the data of 37 subjects. The thirty-seven experiment sessions include a total of 1,917 dialogue turns (980 by the wizards and 937 by the students). The students tried maximally four different exercises each. Unlike in the first experiment, the time spent on an exercise was not strictly limited. However, since the duration of the experiment session was limited to two hours, some students did not have the opportunity to do all exercises. The fourth exercise was considered a "challenge exercise", and therefore we expected that only some students would attempt it. On average, each student attempted 2.7 exercises (i.e. we have collected a total of 100 exercise-subdialogues).

The dialogues were annotated by the wizard during the experiment and recorded by DiaWOz-II. Any student utterance that represents a proof step was classified by the wizards w.r.t. three dimensions; *correctness* (i.e. *correct*, *partially correct* or *incorrect*), the *step size* (i.e. *appropriate*, *too detailed* or *too coarse-grained*), and *relevance* (i.e. *relevant*, *restricted relevance* or *irrelevant*). Statements from the student which did not represent a contribution to the proof (e.g. meta-comments or questions) were annotated with a placeholder *unknown*. A dialogue excerpt from the

T1: Let R and S be relations in an arbitrary set M. It holds that: $(R \circ S)^{-1} = S^{-1} \circ R^{-1}$. Do the proof interactively with the system.
S1: Help me with the proof of the exercise
`(unknown; unknown; unknown)`
T2: In order to show the equality between two sets, one generally shows that the first set is a subset of the second set and vice versa
S2: $R^{-1} \subset R^{-1}$ `(incorrect; appropriate; irrelevant)`
T3: No, you first have to show: $(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$
S3: $(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$
`(unknown; appropriate; relevant)`
T3: Yes, that is to be shown. Therefore take an element from $(R \circ S)^{-1}$ and show that it is also in $S^{-1} \circ R^{-1}$.
S4: $R \subseteq S^{-1} \circ R^{-1}$
`(incorrect; appropriate; irrelevant)`
T5: No, that is not the right approach. Begin like this: Let $(x,y) \in (R \circ S)^{-1}$.
S5: $(x,y) \in (R \circ S)^{-1}$ `(correct; appropriate; relevant)`
T6: Now try to draw conclusions from this!
S6: $(x,y) \in S^{-1} \circ R^{-1}$
`(correct; too coarse-grained; relevant)`
T6: This cannot be concluded so directly. You need some intermediate steps.

Fig. 2 Dialogue excerpt from the second experiment. Annotations indicate the *correctness*, *granularity* and *relevance* of the student's proof step as judged by the tutor

experiment together with annotations is shown in Fig. 2. In addition to the log-files recorded by DiaWOz-II, screen recordings were made. Furthermore, the participants were encouraged to "think aloud" and they were audio-recorded and filmed. This comprehensive collection of data not only documents the text of the tutorial dialogues, but also allows us to analyse how the participants used the interface and the study material. The resulting corpus exhibits variety in the use of natural language and mathematical style. This variety enabled us – besides studying the task of proof step evaluation, as presented in the next section – to study the influence of the instructions presented to the students on the use of natural language, as illustrated in [9].

3 Main Challenges and Resources for Proof Tutoring

An analysis of our corpora revealed various challenges for the automation of proof tutoring based on flexible student–tutor dialogues. We present some of the main challenges here and point to the resources required by the tutor system to fruitfully address them. The success of mathural dialogue-based proof tutoring depends on:

A. The student's knowledge and his learning abilities.
B. The tutor system's mathural processing and mathural generation capabilities.
C. The tutor system's ability to maintain and manage the dialogue state and the proof under construction.
D. The tutor system's capability to dynamically judge about the proof steps uttered by the student.
E. The tutor system's capabilities to perform its proof step analysis tasks with respect to a dynamically changing tutorial context.
F. The tutor system's capabilities for a fine grained analysis of erroneous proof steps.
G. The didactic strategy for feedback generation employed in the tutor system.
H. The tutor system's capability to flexibly interleave the above tasks.

We now discuss the challenges for the tutor systems, that is, aspects B–H, in more detail. We take an application perspective on student modelling (challenge A) for addressing some of these aspects; therefore those aspects of student modelling relevant for proof tutoring will be discussed within the frame of challenges B–H.

3.1 B: Mathural Processing and Mathural Generation

An essential capability of human tutors in mathematics is their ability to successfully *communicate* with students. This communication process includes the task of processing the student's utterances as well as the generation of feedback understandable by the student. These processing and generation capabilities of the human

tutor thus constitute an essential resource with respect to his success as a maths tutor. Analogously, powerful analysis and generation capabilities are amongst the most important resources required for any proof tutoring system which is based on flexible dialogues.

Processing natural language with embedded mathematical content, however, is a highly challenging task by itself. The research carried out within the DIALOG project on the analysis of such utterances, with interleaved linguistic and mathematical content, is presented in Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring System of this volume). At the proof level, analysing students' input involves the problem of content underspecification and ambiguous formulation. Interestingly, underspecification also occurs in shaped-up textbook proofs [43]. To illustrate proof-step underspecification let us consider the dialogue excerpt in Fig. 3:

T1: Please show : $K((A \cup B) \cap (C \cup D)) = (K(A) \cap K(B)) \cup (K(C) \cap K(D))$
S1: by the deMorgan rule $K((A \cup B) \cap (C \cup D)) = (K(A \cup B) \cup K(C \cup D))$ holds.

Fig. 3 An excerpt from the corpus of the first experiment

In Fig. 3, the proof-step that the utterance *S1* expresses is highly underspecified from a proof construction viewpoint: it is neither mentioned how the assertion is related to the target formula nor how and which deMorgan rule was used. *S1* can be obtained directly from the second deMorgan rule $\forall X, Y.K(X \cap Y) = K(X) \cup K(Y)$ by instantiating X with $(A \cup B)$ and Y with $(C \cup D)$. Alternatively, it could be inferred from **T1** by applying the first deMorgan rule $\forall X, Y.K(X \cup Y) = K(X) \cap K(Y)$ from right to left to the subterms $K(A) \cap K(B)$ and $K(C) \cap K(D)$. Successful proof tutoring requires that the meaning of the student utterance can be sufficiently determined to allow further processing. The capability to differentiate and prioritise between proof construction alternatives as illustrated by our example is thus an important resource of a tutoring system. And as illustrated, mathural processing may involve non-trivial domain-reasoning tasks (here theorem-proving tasks).

The corpora also illustrate the style and logical granularity of human-constructed proofs. The style is mainly declarative, for example, the students declaratively described the conclusions and some (or none) of the premises of their inferences. This is in contrast to the procedural style employed in many proof assistants where proof steps are invoked by calling rules, tactics or methods, i.e. some proof refinement procedures. The hypothesis that assertion level reasoning [22] plays an essential role in this context has been confirmed. The fact that assertion level reasoning may be highly underspecified in human-constructed proofs, however, is a novel finding [3].

In this chapter we will not further discuss the processing and generation of mathural language. For the processing of utterances with embedded mathematical content with respect to input interpretation and dialogue modelling we refer to Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring System of this volume). In the following we assume that the meaning of a student utterance can always be

successfully determined by the mathural processing resources available to the tutor system.[3]

3.2 C: Dialogue State and Proof Management

The successive dialogue moves performed by student and tutor form a dialogue state which is the context for the analysis of further moves. Part of this dialogue state is an incrementally developing partial proof object which is (hopefully) shared between the student and the tutor. It represents the status of the proof under development by the student at the given point in the dialogue. The maintenance and manipulation of such dynamically changing proof objects thus have to be realised in a proof tutoring system. Ideally the formalised proof objects in a tutor system closely match the mental proof objects as shared by students and human tutors. In particular, to support cognitively adequate proof step evaluation they should not differ significantly with respect to the underlying logical calculus and the granularity of the proof steps.

3.3 D: Proof Step Evaluation

A human tutor who has understood a proof step utterance of his student will subsequently analyse it in the given tutorial context. A main task thereby is to evaluate the *correctness*, *granularity* and the *relevance* of the student proof step in the given tutorial context. Let us neglect the tutorial context for the moment and concentrate, for better understanding, solely on the pure logical dimension of the problem. This pure logical dimension will be illustrated using the artificially simplified example in Fig. 4.

Proof State	**Some Student Utterances**
(A1) $A \wedge B$.	(a) From the assertions follows D.
(A2) $A \Rightarrow C$.	(b) B holds.
(A3) $C \Rightarrow D$.	(c) It is sufficient to show D.
(A4) $F \Rightarrow B$.	(d) We show E.
(G) $D \vee E$.	

Fig. 4 PSE example scenario: (A1)–(A4) are assertions that have been introduced in the discourse and that are available to prove the proof goal (G). (a)–(d) are examples for possible proof step directives of the student in this proof situation

Correctness analysis requires that the domain reasoner can represent, reconstruct and validate the uttered proof step (including all the justifications used by the student) within the domain reasoner's representation of the proof state. Consider, for

[3] Note that in practice we can support mathural processing with the help of clarification subdialogues or by appropriately restricting the flexibility of the mathural language.

instance, utterance (a) in Fig. 4: Verification of the soundness of this utterance boils down to adding D as a new assertion to the proof state and to proving that *(P1)* $(A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash D$. Solving this proof task confirms the logical soundness of utterance (a). If further explicit justifications are provided in the student's utterance (e.g. a proof rule) then we have to take them into consideration and, for example, prove *(P1)* modulo these additional constraints.

Granularity evaluation requires analysing the "complexity" or "size" of proofs instead of asking for the mere existence of proofs. For instance, evaluating utterance (a) above boils down to judging the complexity of the generated proof task *(P1)*. Let us, for example, use Gentzen's natural deduction (ND) calculus as the proof system $\vdash$. As a first and naive logical granularity measure, we may determine the number of $\vdash$-steps in the smallest $\vdash$-proof of the proof task for the proof step utterance in question; this number is taken as the argumentative complexity of the uttered proof step. For example, the smallest ND proof for utterance (a) has "3" proof steps: we need one "Conjunction–Elimination" step to extract A from $A \wedge B$, one "Modus Ponens" step to obtain B from A and $A \Rightarrow B$, and another "Modus Ponens" step to obtain C from B and $B \Rightarrow C$. On the other hand, the smallest ND proof for utterance (b) requires only "1" step: B follows from assertion $A \wedge B$ by "Conjunction–Elimination". If we now fix a threshold that tries to capture, in this sense, the "maximally acceptable size of an argumentation" then we can distinguish between proof steps whose granularity is acceptable and those which are not. This threshold may be treated as a parameter determined by the tutorial setting. However, as we will further discuss in Sect. 4, using ND calculus together with a naive proof step counting is generally insufficient to solve the granularity challenge. More advanced approaches are needed.

Relevance asks questions about the usefulness and importance of a proof step with respect to the original proof task. For instance, in utterance (c) the proof goal $D \vee E$ is refined to the new proof goal D using backward reasoning, i.e. the previously open goal $D \vee E$ is closed and justified by a new goal. Answering the logical relevance question in this case requires to check whether a proof can still be generated in the new proof situation. In our case, the task is thus identical to proof task *(P1)*. A backward proof step that is not relevant according to this criterion is (d) since it reduces to the proof task: *(P2)* $(A \wedge B), (A \Rightarrow C), (C \Rightarrow D), (F \Rightarrow B) \vdash E$ for which no proof can be generated. Thus, (d) is a sound refinement step that is not relevant.

3.4 E: Tutorial Context

Dynamic proof step evaluation enables tutoring in the spirit of didactic constructivism [25] (i.e. allowing the student to explore rather than expect him to follow a prescribed solution path). Dynamic proof step evaluation poses already a non-trivial challenge to mathematical domain reasoning if we assume a static tutorial context. In practice, however, the tutorial context dynamically changes. This tutorial

context comprises the dynamically changing knowledge and experience of the student, the possibly dynamically changing teaching goal and strategy, and the dynamically changing knowledge of the teacher about the student's dynamically changing capabilities. Incorporating this dynamically changing context information poses an additional challenge to the tutor system's proof step evaluation mechanism since the system needs to adapt its analysis to both the tutorial model and the student model.

3.5 F: Failure Analysis

Context sensitive proof step evaluation supports the separation of acceptable from unacceptable student proof steps. In case of acceptable proof steps the student will be encouraged to continue his proof. More challenging is to compute and present useful feedback also in the case of unacceptable proof steps, that is, proof steps which are erroneous. Standard tutoring systems typically rely on information provided in advance by the author of teaching materials. Since we are in a setting where solutions are determined on the fly, we face the issue whether solution proofs can be dynamically annotated with information on the reason for failure. Such additional information provides important input for the generation of didactic useful feedback. In order to obtain such additional information of the reasons for failure the tutoring system needs to dynamically solve further analysis tasks in the domain reasoner.

3.6 G: Didactic Strategies, Feedback Generation and Hinting

The tutor can decide to offer a hint to the student in a number of situations, for example after repeated student errors, a long period of silence or a direct request. Given information about the proof step such as correctness, granularity and relevance as well as information about the student – encoded in the student model – the tutor should react in a way that optimises the progress of the student. In general, the behaviour of the tutor is encoded in a teaching strategy. *Socratic teaching* strategies that focus on posing questions to that student, not answers, have shown to be more effective than simply presenting the student with concrete parts of the solution.

3.7 H: Flexible Dialogue Modelling

Human maths tutors usually show impressive skills with respect to (at least) all of the aspects above. These tutoring skills – typically they are acquired in special training courses – are important resources limiting the tutor's capabilities for effective proof tutoring. These skills are consequently also important resources for an automated proof tutor system. It requires modules addressing these skills and these modules need to interact in a suitable way. Controlling the overall dialogue, invoking these modules, combining their results and determining appropriate dialogue moves

is the task of the dialogue manager. Human tutors are generally capable of flexibly applying and interleaving their tutoring skills. This calls for flexible approaches and flexible architectures for dialogue management to convey this flexibility of human tutors to the tutor system. An example for an interleaving of skills has been hinted at before: in order to disambiguate a content-underspecified proof step utterance of the user, mathural processing may want to consult the proof manager and proof step evaluation in order to rule out incorrect readings. Details on the dialog management architecture are found in Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring System of this volume).

4 Dynamic Proof Step Evaluation with ΩMEGA

A main focus in the DIALOG project has been on proof step evaluation. We have already argued that dynamic, context sensitive proof step evaluation requires support from a sophisticated and ideally cognitively adequate mathematical domain reasoner.[4] The ΩMEGA system, with its various support tools for human-oriented, abstract level proof representation and proof construction, has therefore been chosen as the domain reasoner of choice in the DIALOG project (see Wolska *et al.*, Linguistic Processing in a Mathematics Tutoring System of this volume).

4.1 Proof Management, Correctness Analysis and Content Underspecification

In the DIALOG context ΩMEGA is used to (i) represent the mathematical theory in which the proof exercise is carried out, that is definitions, axioms and theorems of a certain mathematical domain, (ii) to represent the ongoing proof attempts of the student, in particular the management of ambiguous proof states resulting from underspecified or ambiguous proof steps the student enters, (iii) to maintain the mathematical knowledge the student is allowed to use and to react to changes of this knowledge and (iv) to reconstruct intermediate steps necessary to verify the correctness of a step entered by the student, thereby also resolving ambiguity and underspecification.

Proofs are represented in ΩMEGA's proof data structure (PDS) which allows the shared representation of several (ongoing) proof variants [4]. The PDS is a collection of proof trees (with nodes as multi-conclusion sequents), which can be linked to one another (in order to express the dependency of one proof on another one whose proof task is treated as a lemma).

---

[4] There is a discrepancy between the level of argumentation in mathematics and the calculus level in contemporary automated theorem proving. We argue that cognitively motivated domain reasoning, such as reasoning at the assertion level, can overcome the limitations of theorem proving with commonly used calculi such as resolution or natural deduction calculi (cf. [11]).

These reconstructed proofs serve as the basis for further analysis of the students' proof steps w.r.t. granularity and relevance. Thereby, our analysis components take advantage of ΩMEGA's abstract proof representation at the assertion level. We now sketch some of the project achievements.

As the basis for proof step evaluation, each proof step proposed by the user is reconstructed in ΩMEGA. As explained before for the example student utterance **S1** in Fig. 3, even most ordinary human proof steps can generally include a number of tacit intermediate steps, which become apparent when modelling these proof steps in a rigorous formal system. Therefore, the reconstruction generally requires proof search in order to determine the different (correct) readings of the student proof step.

The assessment module we have realised as part of the ΩMEGA system maintains an assertion level proof object that represents the current state of the proof under construction, which can include several proof alternatives in the case of underspecified, that is, insufficiently precise, proof step utterances by the student causing ambiguities (cf. [5, 15]). For each proof step uttered by the student, the module uses a *depth-limited breadth-first search* (with pruning of superfluous branches) to expand the given proof state to all possible successor states up to that depth. From these, those successor states that match the given utterance w.r.t. to some filter function (analysing whether a successor state is a possible reading of the student proof step) are selected. Thus, we have combined the resolving of *content underspecification and the verification of the correctness of a proof step* as a joint task in our solution in ΩMEGA. If a student proof step matches with a step in one of the possible assertion level proofs (expanding the current proof state to a certain depth) as generated by ΩMEGA, then it is considered as correct. The matcher and intermediate steps in the ΩMEGA proof object not addressed by the student are then the formally relevant content that was left unspecified in the student utterance.

This way we obtain, modulo our filter function, assertion level counterparts to all possible interpretations of correct student proof steps. If for a given utterance, no matching successor state can be reached, the utterance is considered as incorrect.

In [11] we report on a case study in which we applied our ΩMEGA based assessment module with a depth-limit of four assertion level steps to 17 dialogues from the second DIALOG corpus; these 17 dialogues contain a total of 147 proof steps. All the steps within a dialogue were passed to the assessment module sequentially until a step that is labelled as correct cannot be verified, in which case we move on to the next dialogue. This way, we correctly classify 141 out of the 147 steps (95.9%) as correct or wrong. Among the remaining six steps are three where the verification fails, and further three remain untouched.

This experiment confirms our decision in the ΩMEGA project to replace the previous ND-based logical core by assertion level reasoning: Using breadth-first proof search for automated proof step analysis in proof tutoring appears unreasonable at first sight (if we employed ND calculus it likely would be, because of the challenging, deep search space; it remains unclear whether strategically guided ND proof search as employed, e.g. in AProS [38] can sufficiently reduce the search). However, we employ assertion level breadth-first search which turns out to be well suited for the given task. This is because in ΩMEGA we obtain more adequate formal counter-

parts of the human proofs as was possible before: Looking at the human level proof steps in our corpus from the perspective of assertion level reasoning our analysis shows that they seldom exceed size three. Interestingly, already a depth limit of just four assertion level steps enables our breadth-first search-based approach to correctly classify 95.9% of the proof steps in the corpus of the experiment.

4.2 Granularity Analysis

An early study within this project on granularity [34–36] investigated the use of proof reconstructions in natural deduction calculi for obtaining a measure for granularity. We investigated the viewpoint outlined in Sect. 3, where the number of steps in a formal deductive system (here, a natural deduction system) is treated as an indicator for granularity. Natural deduction is a self-evident first candidate for modelling human proof steps. We studied two human-oriented calculi, the classic natural deduction calculus by Gentzen [19] and the more recent psychologically-motivated calculus PSYCOP [20]. For our investigation, we made use of the proofs in the experiment corpus of the second Wizard-of-Oz experiment, where each step from the student is annotated with a granularity judgment by the human wizard, which can take one out of three values *too detailed, appropriate* or *too coarse-grained*.

In particular, we related the step size of proofs in the experiment corpus (as indicated by the wizards) to the step size of these two calculi. As reported in [36], large (i.e. *too coarse-grained*) proof steps (as identified by the wizards) corresponded usually to longer sequences of natural deduction inference applications. However, a large gap w.r.t. step size remained between human-generated proofs and natural deduction proofs. A single proof step as it typically occurs in the experiment corpus generally requires numerous deduction steps at the level of natural deduction, which are often of rather technical nature. It became apparent that the sheer number of inference steps in the natural deduction proof reconstructions was an insufficient measure for granularity.

4.3 Learning Granularity Evaluation

The previous studies [36, 34] motivated the investigation of assertion-level proof reconstructions in ΩMEGA as a basis for granularity analysis. Furthermore, the experiments hinted at other criteria as indicators for steps size besides counting the number of calculus-level inference steps required to reconstruct a human-made proof step. Based on the experiment corpus, we have identified a number of potentially granularity-relevant criteria.

Homogeneity: A single human-made step that involves the application of several different mathematical facts is distinguished w.r.t. granularity from a step where only one fact is applied several times.

Verbal Explanation: A human-made step that is accompanied by verbal justification of the argumentation (e.g. the name of a theorem, definition, etc.) is distinguished w.r.t. granularity from a step where only the result (possibly only a formula) is given.

Introduction of Hypotheses or Subgoals: Proof steps which introduce a new hypothesis or new subgoal are given a special status w.r.t. granularity.

Learning Progress: A proof step that involves (one or several) concepts that are known to the student (as recorded by a student model) can be distinguished w.r.t. granularity from proof steps involving (one or several) yet unknown (i.e. to-be-learnt) concepts.

For a given proof step, these criteria can easily be determined from the proof step's assertion-level reconstruction and with the help of a student model. Since, for example, proof reconstruction with ΩMEGA delivers the mathematical assertions employed in the reconstruction process (i.e. facts such as definitions, theorems and lemmata), the question whether these assertions are already known to the user can be answered by a simple lookup in the student model.

However, this leaves the question open in how far each of the criteria contributes to the overall verdict on granularity for that step.

We have developed an approach to learning the relationship between the criteria and the final granularity judgments from an empirical corpus, using standard machine-learning techniques. This allows us to adapt our framework to a particular mathematical domain and the style of a particular human tutor. Thus, we employ two modules for granularity analysis (see Fig. 5); one serves to obtain training instances, from which the associations between granularity criteria and granularity judgements can be learned. This results in a classifier, which is used within a second judgement module to automatically perform granularity judgements.

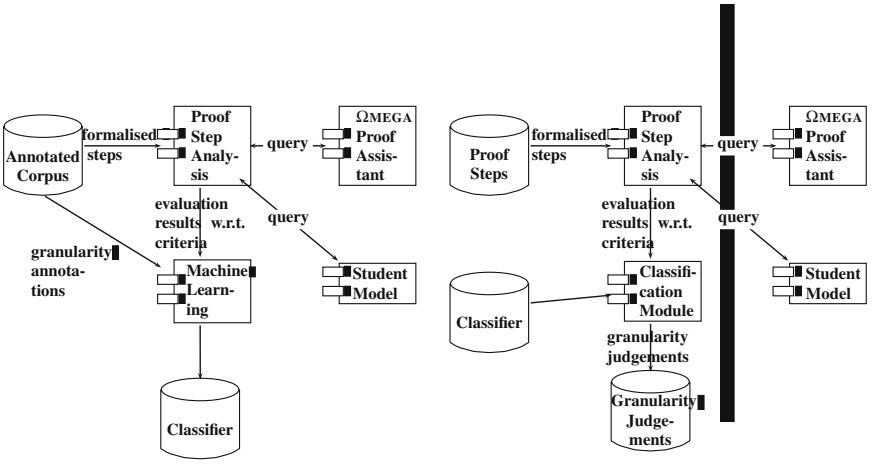


Fig. 5 Training module (*left*) and judgment module (*right*)

Training instances can be constructed from annotated corpora such as in the second experiment. Consider for example the utterance *S5* in Fig. 2, which is the first step in the dialogue the tutor considers correct. This utterance by the student is sent to the proof step analysis module (see Fig. 5), and again handed over to ΩMEGA for proof reconstruction, where it advances the proof state maintained by ΩMEGA by two assertion applications: (i) the (backward) application of the definition of $=$ (such that $(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$ and $S^{-1} \circ R^{-1} \subseteq (R \circ S)^{-1}$ remain to be shown) and (ii) the (backward) application of the definition of $\subseteq$, i.e. in order to show $(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$ it is assumed that $(x, y) \in (R \circ S)^{-1}$ and $(x, y) \in S^{-1} \circ R^{-1}$ remains to be shown. Proof step reconstruction thus delivers the information that two different concepts (the definitions of $=$ and $\subseteq$) were possibly employed by the student in utterance *S5*. The proof step is now analysed with respect to the granularity criteria and a student model (which is updated during the course of the exercise). The results of our evaluation of the criteria for a given proof step are only numeric; they indicate how many assertion-level steps the reconstruction contains (in our running example "2"), how many different concepts the reconstruction involves (again "2"), how many inference steps are unexplained ("2", since the student does not mention the concepts verbally, this would have been for example: "By the definition of equality and the subset relation, . . ."), how many times (if any) new subgoals or hypotheses are introduced (here "1" hypothesis and "3" new subgoals), and how many concepts are new to the learner, according to the student model ("0", if we assume the student is familiar with naive set theory, and in particular equality and subset relation).

For each student step, these results of the analysis are combined with the judgement from the tutor, which is stored in the corpus, and the resulting instance is added to the set of training instances for machine learning. In our running example, the values of the analysis are associated to the verdict "appropriate"; thus they become an example of an *appropriate* step for the machine learning algorithm. However, the evaluation values for the next step *S6* become an example for a *too coarse-grained* student step. The task performed by machine learning is to build a model of these examples (on a given training sample) that allows us to classify further new, yet unseen instances according to their granularity. Like the training module, the judgment module receives student proof steps and analyses them with the help of ΩMEGA and the student model. However, the classifier learnt with the help of the training module now permits automatic granularity judgments.

Currently, we use C5 decision tree learning (see [33] and also [30]) as the learning algorithm. We have also compared this to the performance of other machine learning algorithms on our data, as reported in [37]. One result is that the classifier SMO [29], which implements a support vector machine, achieved a better classification on our sample from the experiment corpus than C5 (see Fig. 6).

However, using decision tree learning (as with C5) has the additional value that the resulting decision trees can easily be interpreted, and thus reveal which of the criteria are relevant to the granularity decisions (w.r.t. the particular corpus and a particular tutor). For example, a case study on a small test set produced the following decision tree depicted in Fig. 7.

	Naive classification	C5	SMO
Mean error	25.6%	13.0%	6.4%
Kappa	0.0	0.65	0.84

Fig. 6 Performance of C5 and SMO on a sample of 47 proof steps from our corpus using ten fold cross validation, compared to naively assigning all proof steps to the majority class *appropriate*

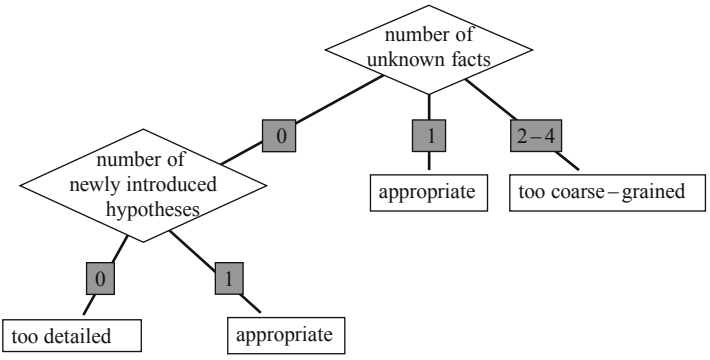


Fig. 7 Example decision tree produced during a case study

In this example, the algorithm has learnt that for the particular "judge" who has trained the system in the case study, the number of previously unknown facts (i.e. facts that the student has not used before) and the number of newly introduced hypotheses are those criteria that explain the judge's behaviour w.r.t. granularity best. However, the criterion of verbosity has been pruned from the tree, which indicates that this criterion is not relevant for the particular training sample. Note that the judge only provides examples and does not need to reflect about granularity criteria or the working of the training module at all.

4.4 Student Modelling

Student modelling is an indispensable ingredient for the analysis of the student's proof attempts and a prerequisite for the generation of tailored feedback. We draw on the experience of the ActiveMath project with various techniques of student modelling and employ a student model based on [28], which is currently used for the LeActiveMath system [20]. For each concept, eight competencies are modelled, namely to *think, argue, model, solve, represent, language, communicate* and *use tools* w.r.t. the concept. The degree of mastery w.r.t. each of these competencies is expressed in four levels: *elementary, simple-conceptual, complex* and *multi-step*. Using this well-established approach to student modelling allows us to embed the techniques of the DIALOG project into the LeActiveMath (and possible other) teaching environments.

The student model is seen as a dynamic component that further contributes to the adaptivity of the system. Competency levels in the student model associated with mathematical facts are updated whenever the mathematical fact is successfully applied, i.e. when it is employed in a student proof step verified by ΩMEGA as correct and categorised as appropriate w.r.t. granularity. Furthermore, each confirmed proof step is turned into a lemma (in case it is not already part of the assertions for the given mathematical domain) and added to the set of available assertions in ΩMEGA, together with a new student model entry. This allows to model common mathematical practice, where previously solved problems become building blocks for subsequent proof construction.

4.5 Further Work

As mentioned, a start has been made to incorporate a student model into granularity evaluation. However, more work is needed to incorporate further tutorial context information into proof step evaluation, in particular, into correctness and relevance evaluation. Relevance has generally only been preliminarily addressed in this project so far. This also applies to fine-grained failure analysis.

5 Didactic Strategies and Dialogue Modelling

The socratic teaching challenge has not been a main research focus of the DIALOG project. However, in close collaboration with our project, Tsovaltzi and Fiedler have studied hint taxonomies [40, 41] and dialogue-adaptive hinting [17].

5.1 Didactic Strategies and Hinting

The approach described in [17] is to dynamically produce hints that fit the needs of the student with regard to the particular proof. Thus, the proof tutor system cannot restrict itself to a repertoire of static hints, associating a student answer with a particular response by the system. Reference [41] defines a multi-dimensional hint taxonomy where each dimension defines a decision point for the associated cognitive function. The domain knowledge can be structured and manipulated for tutoring decision purposes and generation considerations within a tutorial manager. Hint categories abstract from the strict specific domain information and the way it is used in the tutoring, so that it can be replaced for other domains. Thus, the teaching strategy and pedagogical strategy of the proof tutor system can be retained for different domains. More importantly, the discourse management aspects of the dialogue manager can be independently manipulated.

The hint taxonomy [41] was derived with regard to the underlying function of a hint that can be common for different NL realisations. This function is mainly responsible for the educational effect of hints.

5.2 Dialog Modelling

Dialogue systems act as conversational agents, that is, they look at an analysis of an incoming utterance and call on conversational expertise, encoded for instance as a dialogue grammar, to determine an appropriate response in the current dialogue state. A model of dialogue state, containing for example a record of utterances and their analyses, is continually updated to reflect the effect of both system and user utterances. In order to successfully manage and model tutorial dialogue, the dialogue state must be centrally stored and the results of computations by system modules, such as natural language analysis, must be made available. To satisfy these requirements, we have been working within a model characterised by a centrally placed dialogue manager. The dialogue manager maintains the model of dialogue state, facilitates the collaboration of single system submodules and controls top-level execution in the system. By using the current dialogue state and accessing its model of conversational expertise, the dialogue manager is in the position to choose the most appropriate system response.

In developing a suitable model for managing tutorial dialogue we have met a number of challenges: How should we facilitate interleaving the processing carried out by system modules in the analysis of students' utterances? How can we combine the results of each module's analysis into a representation that forms the context of the choice of system response? And what information needs to be modelled at the dialogue level as opposed to the task or tutoring level?

In keeping with our project goal of flexible tutorial dialogues on mathematical proofs, we have been continually developing a demonstrator dialogue system which implements this type of model. It serves as a framework in which single modules can be tested, and this work is presented in Wolska *et al.* (Linguistic Processing in a Mathematics Tutoring system of this volume).

6 Related Work and Conclusion

All existing systems for proof tutoring employ automated theorem-proving techniques for checking, analysing or generating complete proofs. Examples are the proof tutoring systems EPGY [39], ETPS [2], TUTCH [1], INTELLIGENTBOOK [13] and WINKE [14]. If proof checking fails the only feedback these systems return is that the step could not be verified. If the verification succeeds, there is no further analysis whether that proof step is actually relevant to complete the proof or whether it is of appropriate granularity. The ETPS-system is the only system which provides hints about how to continue the proof in case the student gets stuck. An approach that uses information about completed proofs is realised in WHYATLAS [26]. This

system checks a student's proof in two stages: First, it uses domain specific rules to generate different possible proofs for the given conjecture using abductive logic programming [23] and then it compares the found proofs (including those using some didactically motivated buggy rules) to the student's proof and selects the most similar one to provide feedback to the student. Finally, the APROS project [38] uses automated proof search for natural deduction as the basis for the dynamic *Proof Tutor* system. It has been successfully used as a part of the course "Logic & Proofs" at Carnegie Mellon University, which at current time has been attended by more than 2000 students.

Main contributions of our work include empirical experiments that served to pinpoint the particular research challenges evoked by our ambitious dialogue scenario for dynamic proof tutoring. As a result, we determined that proof step evaluation does not only include the aspect of correctness, but also the analysis of granularity and relevance. Besides an elaborate discussion of the various functions relevant for proof tutoring and their overarching architecture, we have examined the analysis of granularity in detail, and developed an adaptive architecture based on the analysis of granularity-related features and machine learning techniques. Our work has resulted in a demonstrator system and prototype implementations, which allowed partial evaluation. Ultimately, such a system should be evaluated regarding its benefits for students' learning performance. However, this is still future work, since it already presupposes a high degree of maturity of the investigated system.

Acknowledgments We are grateful to the reviewers of this paper who provided many useful comments and suggestions. The second author gratefully acknowledges the support from the German National Merit Foundation (Studienstiftung des deutschen Volkes e.V.).

References

1. Abel, A., Chang, B.Y.E., Pfenning, F. Human-readable machine-verifiable proofs for teaching constructive logic. In U. Egly, A. Fiedler, H. Horacek, S. Schmitt, (Eds.), Proceedings of the Workshop on Proof Transformations, Proof Presentations and Complexity of Proofs (PTP'01). Italian: Universitá degli studi di Siena (2001).
2. Andrews, P., Bishop, M., Brown, C., Issar, S., Pfenning, F., Xi, H. Etps: A system to help students write formal proofs. Journal of Automated Reasoning, 32: 75–92 (2004).
3. Autexier, S., Benzmüller, C.E., Fiedler, A., Horacek, H., Vo, B.Q. Assertion-level proof representation with under-specification. Electronic Notes in Theoretical Computer Science, 93:5–23 (2004).
4. Autexier, S., Benzmüller, C., Dietrich, D., Meier, A., Wirth, C.P. A generic modular data structure for proof attempts alternating on ideas and granularity. In M. Kohlhase, (Ed.), Proceedings of the 5th International Conference on Mathematical Knowledge Management (MKM'05) (vol. 3863, pp. 126–142). Heidelberg: Springer, LNAI (2006).
5. Benzmüller, C., Vo, Q. Mathematical domain reasoning tasks in natural language tutorial dialog on proofs. In M. Veloso, S. Kambhampati, (Eds.), Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05) (pp. 516–522). Pittsburgh, Pennsylvania, USA: AAAI Press/The MIT Press (2005).
6. Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Tsovaltzi, D., Vo, B.Q., Wolska, M. Tutorial dialogs on mathematical proofs. In:

Proceedings of IJCAI-03 Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems (pp. 12–22). Acapulco, Mexico (2003).
7. Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Tsovaltzi, D., Vo, B.Q., Wolska, M. A wizard of oz experiment for tutorial dialogues in mathematics. In: Proceedings of AI in Education (AIED 2003) Workshop on Advanced Technologies for Mathematics Education. Sydney, Australia (2003).
8. Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayová, I., Tsovaltzi, D., Vo, B.Q., Wolska, M. Towards a principled approach to tutoring mathematical proofs. In: Proceedings of the Workshop on Expressive Media and Intelligent Tools for Learning, German Conference on AI (KI 2003). Hamburg, Germany (2003).
9. Benzmüller, C., Horacek, H., Lesourd, H., Kruijff-Korbajova, I., Schiller, M., Wolska, M. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In: Proceedings of International Conference on Language Resources and Evaluation (LREC 2006). ELDA, Genoa, Italy (2006).
10. Benzmüller, C., Horacek, H., Lesourd, H., Kruijff-Korbayová, I., Schiller, M., Wolska, M. Diawoz-II – a tool for wizard-of-oz experiments in mathematics. In C. Freksa, M. Kohlhase, K. Schill, (Eds.), KI 2006: Advances in Artificial Intelligence. 29th Annual German Conference on AI (vol. 4314). Heidelberg: Springer, LNAI (2006).
11. Benzmüller, C., Dietrich, D., Schiller, M., Autexier, S. Deep inference for automated proof tutoring? In J. Hertzberg, M. Beetz, R. Englert, (Eds.), KI, Springer, Lecture Notes in Computer Science (vol. 4667, pp. 435–439). New York: Springer (2007).
12. Benzmüller, C., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Wolska, M. Natural language dialog with a tutor system for mathematical proofs. In R. Lu, J. Siekmann, C. Ullrich, (Eds.), Cognitive Systems (vol. 4429). New York: Springer, LNAI (2007).
13. Billingsley, W., Robinson, P. Student proof exercises using mathtiles and isabelle/hol in an intelligent book. Journal of Automated Reasoning, 39:181–218 (2007).
14. D'Agostino, M., Endriss, U. Winke: A proof assistant for teaching logic. In: Proceedings of the First International Workshop on Labelled Deduction. (1998).
15. Dietrich, D., Buckley, M. Verification of proof steps for tutoring mathematical proofs. In R. Luckin, K.R. Koedinger, J. Greer, (Eds.), Proceedings of the 13th International Conference on Artificial Intelligence in Education (vol. 158, pp. 560–562). Los Angeles, USA: IOS Press (2007).
16. Melis, E., Siekmann, J. Activemath: An intelligent tutoring system for mathematics. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L. Zadeh, (Eds.), Seventh International Conference 'Artificial Intelligence and Soft Computing' (ICAISC) (vol. 3070, pp. 91–101). Heidelberg: Springer-Verlag, LNAI (2004).
17. Fiedler, A., Tsovaltzi, D. Domain-knowledge manipulation for dialogue-adaptive hinting. In: C.K. Looi, G. McCalla, B. Bredeweg, J. Breuker, (Eds.), Artificial Intelligence in Education — Supporting Learning through Intelligent and Socially Informed Technology (pp. 801–803). IOS Press, no. 125 in Frontiers in Artificial Intelligence and Applications (2005).
18. Fiedler, A., Gabsdil, M., Horacek, H. A tool for supporting progressive refinement of wizard-of-oz experiments in natural language. In J.C. Lester, R.M. Vicari, F. Paraguaçu, (Eds.), Intelligent Tutoring Systems — 7th International Conference (ITS 2004) (pp. 325–335). New York: Springer, no. 3220 in LNCS (2004).
19. Gentzen, G. Untersuchungen über das logische schließen. Math Zeitschrift, 39:176–210, 405–431 (1934).
20. Goguadze, G., Ullrich, C., Melis, E., Siekmann, J., Gross, C., Morales, R. Leactivemath structure and metadata model. Tech. Rep., Saarland University (2004).
21. Van der Hoeven, J. GNU texmacs: A free, structured, wysiwyg and technical text editor. In D. Flipo, (Ed.), Le document au XXI-ième siècle (vol. 39-40, pp. 39–50). Metz: actes du congrès GUTenberg (2001).
22. Huang, X. Human oriented proof presentation: A reconstructive approach. Phd thesis, Universität des Saarlandes, Saarbrücken, Germany, published by infix, St. Augustin, Germany, Dissertationen zur Künstlichen Intelligenz, Vol. 112, 1996 (1994).

23. Kakas, A., Kowalski, R., Toni, F. The role of abduction in logic programming. In: Handbook of Logic in Artificial Intelligence and Logic Programming. Oxford: Oxford University (1995).
24. Kelley, J.F. An iterative design methodology for user-friendly natural language office information applications. ACM Transactions on Information and System, 2(1):26–41 (1984).
25. Liu, C.H., Matthews, R. Vygotsky's philosophy: Constructivism and its criticisms examined. International Education Journal, 6(3):386–399 (2005).
26. Makatchev, M., Jordan, P., VanLehn, K. Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems. Journal of Automated Reasoning, 32:187–226 (2004).
27. Normand-Assadi, S., Coulange, L., Delozanne, É., Grugeon, B. Linguistic Markers to Improve the Assessment of Students in Mathematics: An Exploratory Study. Intelligent Tutoring Systems (pp. 380–389). New York: Springer (2004).
28. OECD Learning for tomorrow's world – first results from PISA 2003. OECD Publishing (2004).
29. Platt, J. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, A. Smola, (Eds.), Advances in Kernel Methods – Support Vector Learning (pp. 185–208). Cambridge, MA: MIT Press (1998).
30. Quinlan, J.R. C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann (1993).
31. Rips, L.J. The Psychology of Proof: Deductive Reasoning in Human Thinking. Cambridge, MA: MIT Press (1994).
32. Rosé, C.P., Moore, J.D., VanLehn, K., Allbritton, D. A comparative evaluation of socratic versus didactic tutoring. In: Proceedings of Cognitive Sciences Society (2001).
33. RuleQuest Research. Data mining tools see5 and c5.0. `http://www.rulequest.com/see5-info.html` (2007).
34. Schiller, M. Mechanizing proof step evaluation for mathematics tutoring – the case of granularity. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany (2005).
35. Schiller, M., Benzmüller, C. Granularity judgments in proof tutoring. In: Poster papers at KI 2006: Advances in Artificial Intelligence: 29th Annual German Conference on AI, Bremen, Germany (2006).
36. Schiller, M., Benzmüller, C., de Veire, A.V. Judging granularity for automated mathematics teaching. In: LPAR 2006 Short Papers Proceedings, Phnom Pehn, Cambodia, `http://www.lix.polytechnique.fr/ hermann/LPAR2006/short/submi ssion_147.pdf` (2006).
37. Schiller, M., Dietrich, D., Benzmüller, C. Proof step analysis for proof tutoring – a learning approach to granularity. Teaching Mathematics and Computer Science, 6(2):325–343 (2008).
38. Sieg, W. The apros project: Strategic thinking and computational logic. Logic Journal of IGPL, 15(4):359–368 (2007).
39. Sommer, R., Nickols, G. A proof environment for teaching mathematics. Journal of Automated Reasoning, 32:227–258 (2004).
40. Tsovaltzi, D., Fiedler, A. Human-adaptive determination of natural language hints. In C. Reed, (Ed.), Proceedings of IJCAI-05 Workshop on Computational Models of Natural Argument (CMNA) (pp. 84–88). Edinburgh, UK (2005).
41. Tsovaltzi, D., Fiedler, A., Horacek, H. A multi-dimensional taxonomy for automating hinting. In J.C. Lester, R.M. Vicari, F. Paraguaçu, (Eds.), Intelligent Tutoring Systems — 7th International Conference (ITS 2004) (pp. 772–781). Berlin: Springer, no. 3220 in LNCS (2004).
42. Wolska, M., Vo, B.Q., Tsovaltzi, D., Kruijff-Korbayová, I., Karagjosova, E., Horacek, H., Gabsdil, M., Fiedler, A., Benzmüller, C. An annotated corpus of tutorial dialogs on mathematical theorem proving. In: Proceedings of International Conference on Language Resources and Evaluation (LREC 2004). ELDA, Lisbon, Portugal (2004).
43. Zinn, C. Supporting the formal verification of mathematical texts. Journal of Applied Logic, 4(4):592–621 (2006).

Part III
Resource-Adaptive Rationality in Machines

Comparison of Machine Learning Techniques for Bayesian Networks for User-Adaptive Systems

Frank Wittig

1 Introduction: Bayesian Networks in User-Adaptive Systems

During the last decade, Bayesian networks (BNs) have been one of the major research topics in the AI community in the area of reasoning under uncertainty. The READY project has been one of the forerunners along these lines – in particular regarding the application of BNs in the context of user modeling/user-adaptive systems (UASs) over the whole period of the collaborative research program 378.

Right from the beginning, BNs have served in the READY prototypes as the core reasoning paradigm. In the early project phase, the BNs have been modeled manually on the basis of theoretical considerations related to cognitive aspects of the user models. This applies in particular to the amount of time pressure and cognitive load, that users are suffering form while they interact with a dialog system [21]. In a second phase, the focus of the research shifted to dynamic BNs that enabled the READY prototypes to take into account dynamic changes related to the user models (UMs) [20]. Finally and consequently, machine learning (ML) techniques for BNs became a central research topic [24]. With such an approach, data, that has been collected in several psychological experiments, could be exploited explicitly for the construction of empirically grounded user models that have been used by the READY prototypes.

Meanwhile, BNs have become a widely established tool in the software industry. Successful real-world applications cover a large spectrum, ranging from spam filters in email clients to the forecasting of customers' buying behavior in the context of customer relationship management or in the retail industry.

The present chapter focuses on the application of previously existing and the development of new BN learning methods that are able to deal with or that can exploit the characteristics of domains of UASs. Previously, BNs used by UASs have typically been specified manually—on the basis of theoretical considerations (of experts). It seems to be a promising approach to exploit the interaction data that

F. Wittig (✉)
SAP AG, Neue Bahnhofstr. 21, D-66386 St. Ingbert, Germany
e-mail: frank.wittig@sap.com

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_14,

can be collected during the systems' use through the application of ML methods in the design and maintenance phases. To this end, new BN learning and adaptation methods that have been developed as part of the READY project are presented, which jointly aim to address adequately the characteristics and demands of the user modeling context during the learning and adaptation processes. These methods have been evaluated in comparative empirical studies relative to alternative existing standard BN learning procedures.

BNs have become one of the most important tools for representing user models in UASs that have to deal with uncertainty. As a reminder of the basic concepts: A BN represents a joint probability distribution over random variables. It consists of two parts: (a) the structure, a DAG to represent the conditional (in-)dependencies between the variables and (b) conditional probability tables (CPTs) that are associated with the links in the DAG. They contain the conditional probabilities of the variables' states conditioned on the combination of their parents' states. A formal definition and the notation used in this chapter is given in the Appendix.

BNs exhibit properties that make them well suited in a wide range of application scenarios in the user modeling context:

- Modeling and reasoning with uncertainty,
- Reasoning about arbitrary sets of variables,
- Extensibility to influence diagrams,
- Modeling temporal aspects using dynamic BNs,
- Interpretability by causal interpretation of links,
- Exploitation of expert knowledge,
- Extensibility to probabilistic relational models and other object-oriented approaches, and
- Availability of ML techniques.

2 A Framework for Learning Bayesian Networks for User-Adaptive Systems

In the READY project, we developed an integrative conceptualization for learning BNs for UASs. After briefly discussing the list of research questions, we will present this framework in terms of major dimensions that have to be taken into consideration here.

2.1 Machine Learning in User-Adaptive Systems

In particular, when applying ML techniques in the area of UASs we identified the following issues that have to be addressed (see [23] for a detailed discussion of a subset of these issues):

- *How can user models be learned on the basis of sparse training data?*
- *How can large inter-individual differences between users be recognized and represented adequately in the user models?*
- *How can changes of the users' interest and characteristics over time be recognized and modeled?*
- *How can the interpretability of the learned user models be ensured/improved?*
- *How can available prior knowledge about the users be exploited during the learning process?*
- *How can different types of training data be exploited jointly for learning user models?*
- *How can causal relationships in user models be determined using machine leaning techniques?*
- *How can the "Overfitting" phenomenon be avoided/limited?*

Some questions, such as limiting overfitting and ensuring interpretability, are of quite general interest in ML research, but nevertheless they are of particular increased importance when UMs are to be learned. All aspect of research on ML (for BNs) in the READY project has been strongly aiming at providing a solution to some of these questions.

2.2 Learning Bayesian Networks for User-Adaptive Systems

Figure 1 presents an overview of a general conceptualization for learning BNs in UASs. It has been the basis of the research done along these lines in the READY project. We will discuss several crucial aspects of this framework by a consideration of important dimensions.

2.2.1 Learning Offline (Batch) and Learning Online (Adaptation)

During the offline phase, general UMs are learned on the basis of data from other previous system users or data acquired by user studies. These models are in turn used as a starting point for the interaction with a particular new user. The initial general model is then adapted to the individual current user and can be saved after the interaction for future use when this particular user will interact the next time with the system. In such a situation, this individual model can be retrieved from the model base and thus, there is no further need to start with the general model, yielding probably a better adaptation right from the beginning. Note that the offline learning procedure may also yield parametric information on how to adapt to individual users. The general idea behind this approach is that different parts of the learned UM need different ways of adaptation, i.e., some parts need faster adaptation to an individual user than others. Details on this and a comparison of alternative methods of adaptation to individual users will be discussed in a following section.

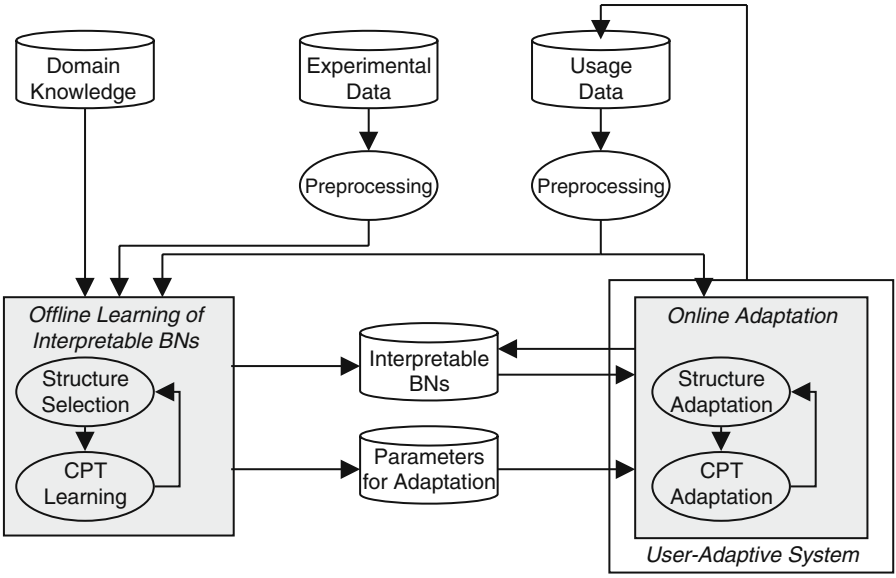


Fig. 1 A framework for learning Bayesian Network user models

2.2.2 Exploiting Experimental and Usage Data for Learning

Two further dimensions concern the kind or type of data that is available. In principle, we distinguish between (a) experimental data and (b) usage data (see upper part of Fig. 1). Experimental data is collected in controlled environments just as done in psychological experiments. Usage data is collected during the real interaction between users and the system. Obviously, these two types differ characteristically: Usage data often includes more missing data and rare situations are underrepresented in such data sets, while experimental data mostly does not represent the "reality". Often, a combination of both types occur. Because of our offline/online approach we can handle this problem for example by learning a general model on the basis of experimental data and then adapting it using usage data of the individual user.

2.2.3 Learning Probabilities and Structure

Since BNs consist of two components, the learning and adaptation tasks are also two-dimensional: (a) learning the conditional probabilities (CPTs) and (b) learning the BNs' structures. For both partial tasks there exist a number of standard algorithms (see [7] for an overview). In the UM context, we often have to deal with sparse data but on the other side in most cases we have additional domain knowledge available. This is reflected in our approach by introducing such knowledge into the learning procedures to improve the results, especially when the data is indeed sparse (see upper left part of Fig. 1). When learning structures, background knowledge can

be incorporated by specifying "starting" structures manually that reflect the basic assumptions that one makes in the domain. This bipartite character of the learning task is reflected in our new techniques we developed for the adaptation of initially learned BN UMs.

2.2.4 Learning Interpretable Bayesian Network User Models

As already discussed, an important point made by many researchers in the user modeling community is the interpretability and transparency of the models. The issue of interpretability is therefore also an integral part of all aspects of our approach. We try to ensure or at least improve the interpretability of the finally learned models, e.g., by respecting whatever background information that is a priori available and that can be introduced into and exploited within the learning process.

2.3 Learning Bayesian Network User Models in the READY Project

In particular, within the presented framework, we achieved the following concrete research results of the READY project related to ML techniques for BNs:

1. *Learning interpretable tables of conditional probabilities using qualitative constraints* [25]
2. *Differential adaptation of conditional probabilities to take into account individual differences between users* [14]
3. *Structural adaptation of BN user models with meta networks*

In this chapter we will focus on the discussion of the latter method itself and interesting empirical results related to its application in the UAS context; regarding the first two, we refer to the previously published articles [25, 14].

The development of these new algorithms enabled the more general research results with regard to the particular domain of the READY project:

1. *It has been shown that it is possible to recognize a user's cognitive resource limitations using learned dynamic BNs on the basis of symptoms of the user's speech.*
2. *Empirically grounded adaptation of the presentation of instructions in a resource-adaptive dialog system using a learned BN, with the goal of avoiding errors and increasing the efficiency of task execution.*

Figure 2 shows the results of an empirical study, whose goal has been to infer the experimental condition of subjects (under time pressure (yes/no)/additional navigational task (yes/no)) based on symptoms of their natural language utterances such as pauses and false starts. For this recognition task we applied learned dynamic BNs. The results of this example scenario clearly show that the experimental condition could successfully be determined with our approach. In sum, as more and more observations (utterances of the experimental subject) become available, the

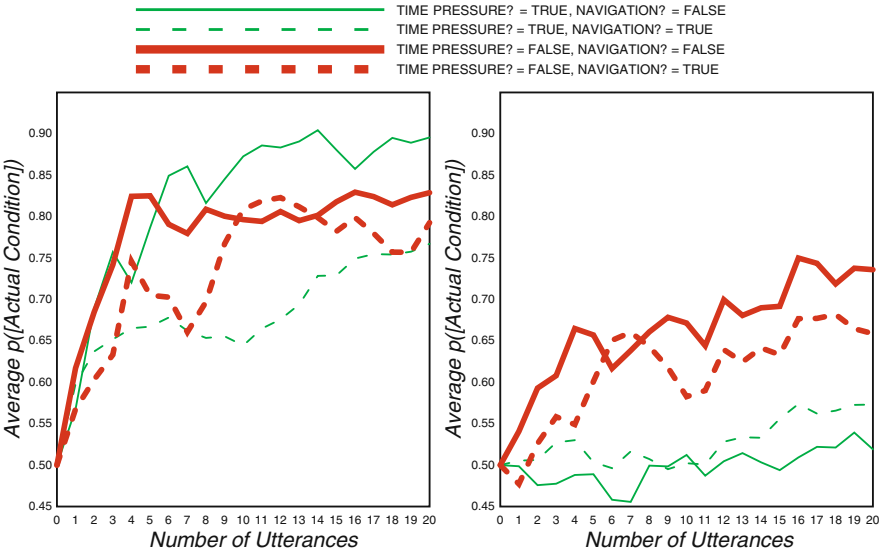


Fig. 2 Recognition results of learned dynamic Bayesian Networks taken from [17]

recognition accuracy increases. A detailed description of the experiment as well as the procedure and results of the empirical study can be found in [17].

Figure 3 shows an influence diagram based on a learned BN. Here again, empirical data gathered within a psychological experiment has been used to learn the BN,

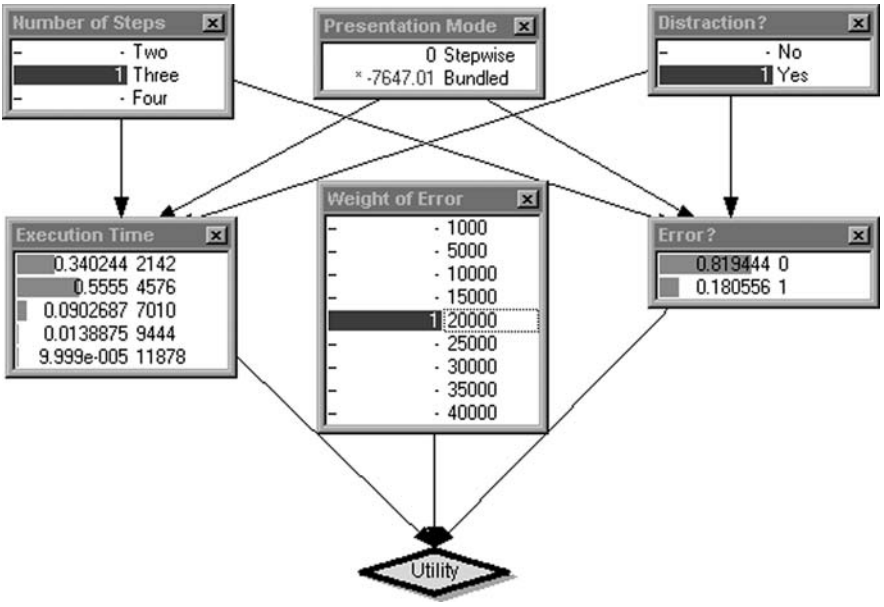


Fig. 3 Influence diagram for adaptation decisions based on learned Bayesian network

that has then been extended to an influence diagram, that could then in turn serve as the core component of a decision engine of a user-adaptive dialog system. In the example scenario, it decides whether a sequence of instructions to the user have to be presented in a "stepwise" or "bundled" mode, taking a trade-off between number of errors and execution time into account. Details can be found in [13].

3 The Structural View: An Evaluation of Bayesian Networks User Model Learning

After we gave an overview of our research on learning BNs for UASs, we will address the hybrid nature of BN user models in more detail. Previously, we have analyzed the learning and adaptation of CPTs in a BN with a fixed structure [25, 14]. Here, we will take a look at several combined approaches that take the structural part of the BN learning task into account.

So far, structural issues have played a minor role in the application of ML techniques for BNs in the user modeling context. Only a few projects exist that have gone beyond the CPT learning/adaptation task [18, 11]. Table 1 shows an overview of relevant research projects. Some reasons for this tendency may be that (a) in many cases it is sufficient to learn the conditional probabilities to receive adequate predictions, (b) it is relatively easy to specify a useful structure based on the causal interpretation of the BNs' links, and (c) structural learning algorithms are too complex for many application scenarios.

Table 1 User-adaptive systems that use machine learning for Bayesian networks

System	Domain	Batch-learning	Adaptation
Albrecht et al. [1]	MUD Games	CPTs	–
Billsus and Pazzani [2]	personalized news	CPTs	CPTs
Lau and Horvitz [15]	WWW search	CPTs	–
Horvitz et al. [10]	Office-alarms	CPTs & struct.	–
Nicholson et al. [18]	ITS	CPTs & struct.	–
Bohnenberger et al. [3]	Dialog	CPTs & struct.	CPTs & struct.

On the other hand, from a knowledge discovery point of view, it is worthwhile to apply structural learning and adaptation methods to improve our understanding of the domain under consideration.

By taking the structural component of BNs into account, we increase the complexity of the space of possibilities. There are more distinctions that have to be considered within the application of such combined user model learning approaches; for example, the structure can be learned for users is general while the CPTs of the same user model are learned for each user individually.

3.1 Combined Learning Approaches

The following combined BN UM learning approaches will be compared. We will make the distinction between parts of the UM that are acquired in "short-term"

vs. a "long-term" way. More concretely, short-time learning is based on the last k observations at a given time in the learning procedure, while long-term learning is based either on all of the currently available data or at least on the data that have become available during the adaptation phase.

- *Long-term individual user model:* With this type of learning, the whole BN is learned on the basis of all of the (adaptation) data that have become available since the interaction started. It is thus an implementation of the purely individual approach. It can be seen as a baseline in that it exploits all available data of the current individual user without taking into account the data of other users.
- *Long-term adaptive with fixed structure/aHugin:* This method uses the AHUGIN method to adapt the CPTs [19] without any change in the structure of the BN. It is of particular interest to see to what extent it is possible to replace possible structural adaptations of the user model (when they are appropriate because of idiosyncrasies of a given user) with adaptations of the conditional probabilities in the CPTs. For example, if optimal structure adaptation for a given user would call for the removal of a given link to node N from node M, a similar effect can be obtained through adaptation of the CPT for node N so that the conditional probabilities reflect the fact that there is no dependence on M. A drawback, relative to structure adaptation, will be potential overfitting because of the unnecessary complexity of the learned model.
- *Short-term individual user model:* A BN–its structure as well as the associated CPTs – is learned anew on the basis of the latest k adaptation cases (the latest adaptation window) for the current individual user. This method represents an extreme point on at least three relevant dimensions: it is based on only a very limited number of observations, and its structure as well as its CPTs are learned within a purely individual approach.
- *Long-term adaptive user model (structural adaptation of BNs with meta-BNs,* SAMBN*):* The SAMBN approach is a method we developed with the user modeling context in mind. It represents a fully adaptive approach, that is, it adapts both the structure and the CPTs of a generally learned UM on the basis of runtime observations of the current user. It maintains an additional data structure – the meta-BN – that is used to reason about the BN UMs structure on a meta-level. That is, it considers reasonable structural alternatives and then chooses the most probable one given the available data. A detailed description of the method is found in the Appendix to this chapter. One of its benefits for the user modeling context is that with its meta-model and the structurally adapted BN UM, it provides a scrutable UM (as far as possible with regard to BNs in general).
- *Short-term individual user model with fixed general structure:* This approach uses the initially (after the offline part) determined general BN UM structure without any further revision and learns the CPTs anew using the k cases of last adaptation window of the current user. In combination with the following approach, this method serves as an indicator for the success of structural adaptation of BNs.

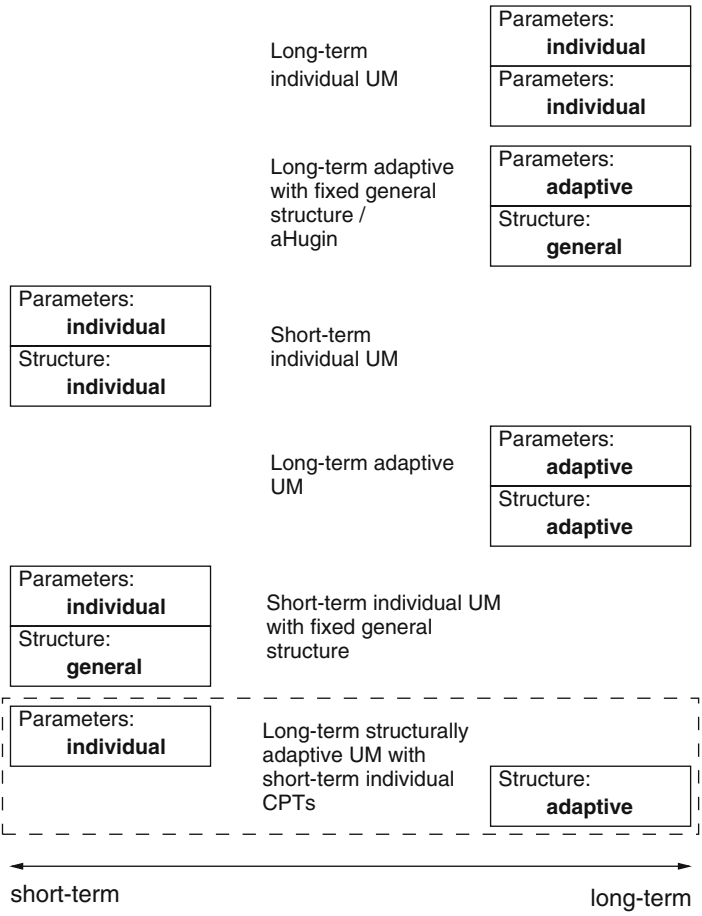


Fig. 4 Combined BN user modeling alternatives

- *Long-term structurally adaptive user model with short-term individual CPTs:* This method adapts the structure according to the structural adaptation algorithm as described in the appendix; but it omits the CPT adaptation and learns the CPTs on the basis of the last adaptation window of the current user. The results of this model represent the potential of the structural adaptation part on its own without influences of the CPTs' adaptation.

Figure 4 characterizes the alternatives according to their type of combined UM learning approach.

3.2 Evaluation Procedure

The procedure to compare the alternative combined BN UM learning approaches is described in the following.

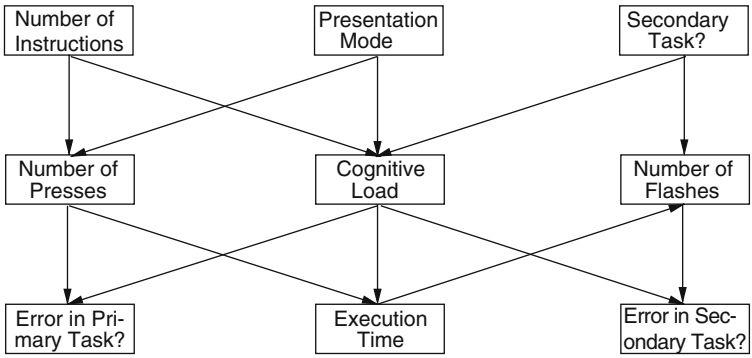


Fig. 5 Structure for experiment 1 used in evaluation

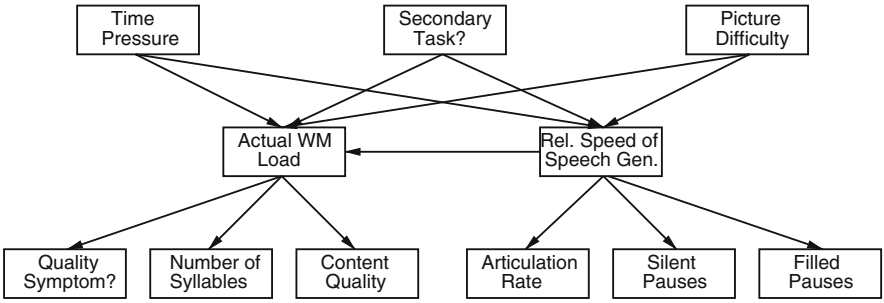


Fig. 6 Structure for experiment 2 used in evaluation

We present results for two different BN structures based on the experiments we performed in our research project as described in [17, 12]. Regarding both BNs, we use the versions shown in Fig. 5 and 6.[1]

We have specified the structures manually and have learned the CPTs on the basis of the available real-world data collected of the experimental subjects.

Both BNs have been used for the generation of five randomly modified BNs in each case. For the first one, there were 3.4 and 2.0 links added and deleted, respectively, for the second structure, the rates account to 1.2 and 3.2. These ten BNs as well as the two original BNs are then in turn used to generate datasets for our analysis.

The evaluation procedure has been as follows: For each alternative learning/adaptation method and example scenario, we start with an offline phase in which a starting BN UM is learned (structure and CPTs) using our empirical datasets. The resulting BN is then adapted to one of the adaptation datasets sampled from the modified BNs – using adaptation windows of k observations. This is done five times, one time for each adaptation dataset. The final results are computed as the average

[1] These complex BN structures are motivated by psychological issues that involve the notion of the user's cognitive load while interacting with the system. We will not go into further detail here and refer to the related publications [3, 17].

values of all five separate runs. The evaluation measure is the normalized log-loss which is a standard measure of performance in density estimation. Essentially, the current BN is scored against the cases of the subsequent adaptation window and the average is computed, see, e.g., [5]. To study the influence of the adaptation windows' size k, we will present the results with different values for this parameter. For those methods that needed the ESS parameter, it has been set to 5.

3.3 Results

Figures 7, 8, 9, 10, 11 and 12 show the results for Experiments 1 and 2 with different sizes k of the adaptation window, respectively. Note that different scales have to be used in the graphs to be able to visualize and subsequently discuss the interesting points.

Overall, the long-term individual UM as the baseline method performed best, as was to be expected. The long-term adaptive UM utilizing the structural adaptation with meta-BNs described in the appendix to this paper was able to outperform the other alternatives clearly – excluding the long-term adaptive UM with a fixed structure, i.e., the AHUGIN method. When the structural part of the method is omitted, i.e., only AHUGIN is applied for the adaptation of the CPTs, we observe a competitive performance. In situations in which more adaptation data are available (larger k values) to determine an adequate modified structure either by learning or adaptation, AHUGIN performed worse in relative comparison, although the absolute performance remained unchanged. All approaches implementing a short-term individual UM show large variation regarding their results. This variation can be

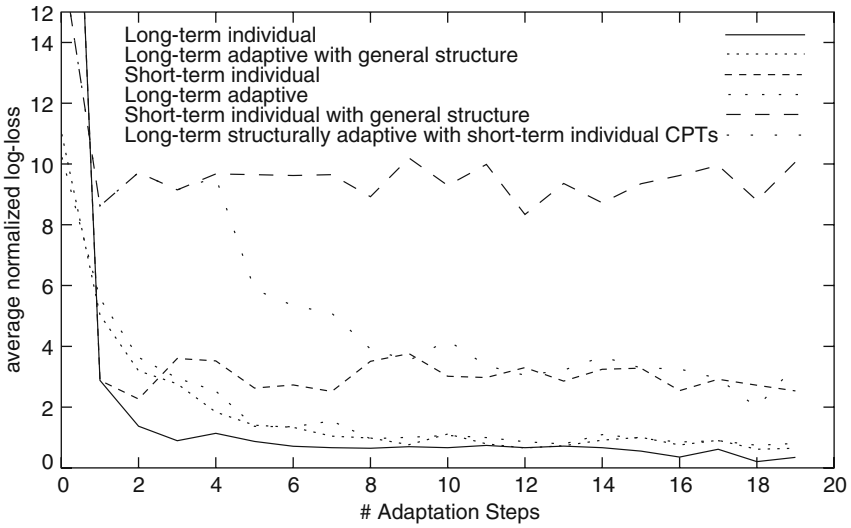


Fig. 7 Results for Experiment 1 with $k = 25$

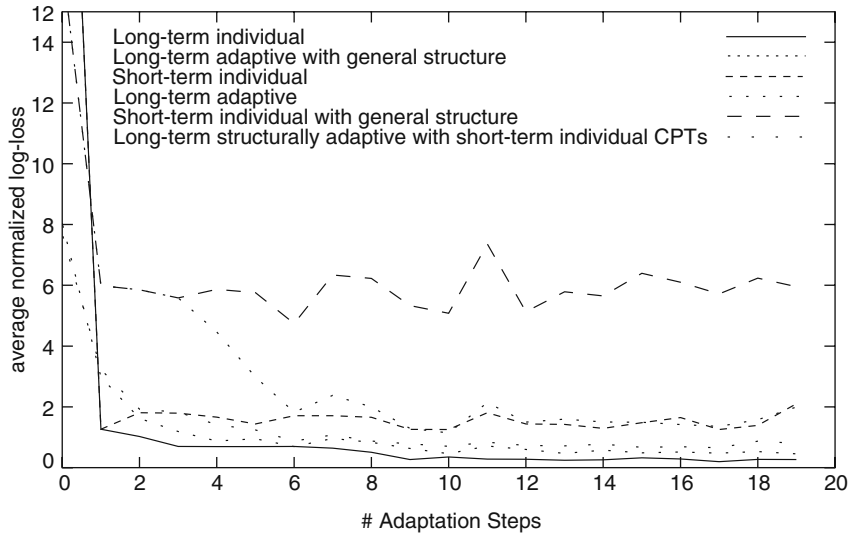


Fig. 8 Results for Experiment 1 with $k = 50$

explained by random fluctuations in the (small) subsequent adaptation sets. This effect was not observed when learning was done on the basis of more adaptation cases (see the results for larger k values). Additionally, as expected, this short-term purely individual method indeed shows the largest improvement in terms of absolute performance when the size of the adaptation window is increased. Considering the

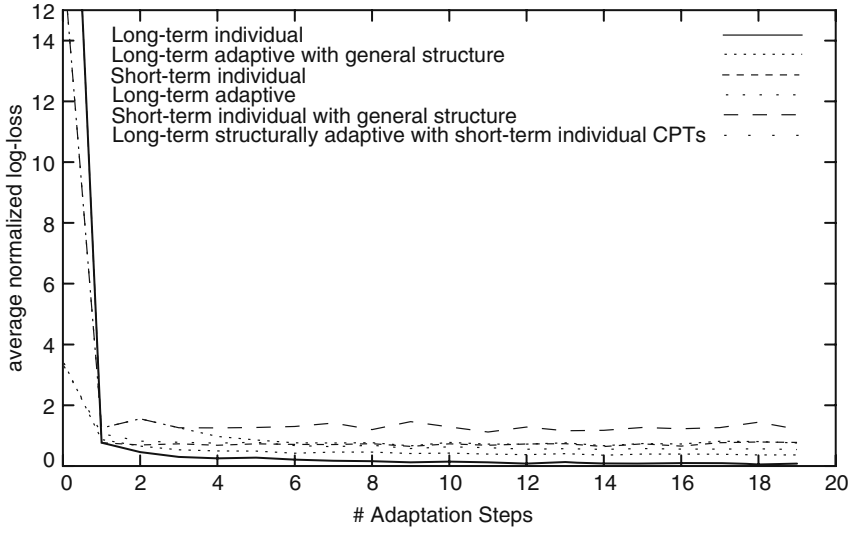


Fig. 9 Results for Experiment 1 with $k = 200$

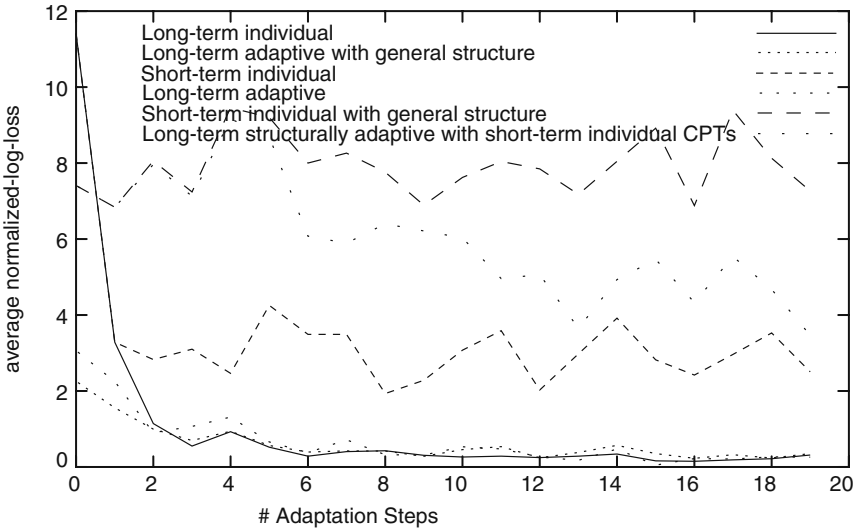


Fig. 10 Results for Experiment 2 with $k = 25$

remaining two procedures, short-term individual/fixed general structure and long-term structurally adaptive/short-term individual CPTs, the results show that structural adaptation by itself (without the AHUGIN adaptation of the CPTs) yields a significant gain in performance.

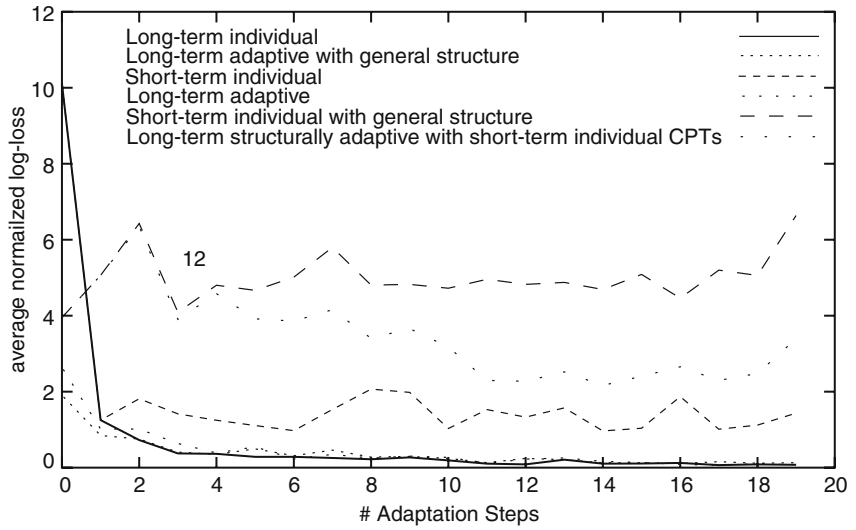


Fig. 11 Results for Experiment 2 with $k = 50$

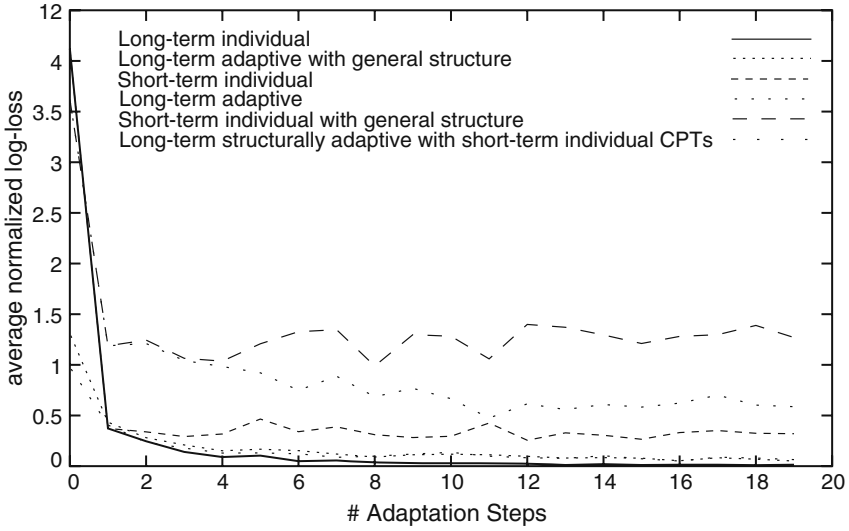


Fig. 12 Results for Experiment 2 with $k = 200$

3.4 Discussion

Table 2 shows a summary of theoretical, empirical, and practical considerations regarding the strength and limitations of the combined BN UM learning approaches.

Altogether, the reasons that speak in favor or against a particular approach have to be considered even more carefully in the light of the demands of the application scenario. The hybrid nature of the BN UMs makes it more difficult to choose the optimal alternative without empirical studies in the domain under consideration. These results should not be seen as a prototypical evaluation; it should be interpreted as a particular example study to discuss some general issues. In other UASs, other combined alternatives may be more promising.

Table 2 Overview of the strength and limitations of the combined alternative approaches

Model type	Theoretical consideration	Empirical results	Practical considerations
Long-term individual	− Concept drift cannot be handled adequately	+ Out-performs all other alternatives	−Complexity increases rapidly over time
Long-term adaptive with fixed structure (AHUGIN)	+ Fewer degrees of freedom compared to structural approaches − Structural differences between users are not reflected in the model	+ Competitive performance in most situations	+ No structural learning/adaptation algorithm has to be applied

Table 2 (continued)

Model type	Theoretical consideration	Empirical results	Practical considerations
Short-term individual	+ UM based on recent data − No longer-term interests considered	− Poor when adaptation window is small	+ Only recent data has to be stored
Short-term individual with fixed structure	− Structural differences between users are not reflected in the model	− Least promising results for all choices of adaptation windows	+ Only recent data has to be stored
Long-term structurally adaptive with short-term individual CPTs	+ Increased scrutability by structural changes	+ Best results of those alternatives that learn individual CPTs	− Structural learning has to be applied at runtime + Only recent data has to be stored + Meta model of the user population
Long-term adaptive (structural adaptation with meta-BNs)	+ Structural representation of individual differences and concept drift	+ Significantly better than long-term structurally adaptive UM with short-term individual CPTs − Quite similar to aHugin	− Structural learning has to be applied at runtime + Only recent data has to be stored + Meta model of the user population

4 Conclusion

The goal of this chapter has been two-fold: (a) to give an overview of the research results of the READY project with regard to learning BNs for UASs and (b) to present an extensive empirical evaluation of (structural) BN learning approaches and a discussion of the results, including a method for structural learning of BN UMs that has been developed by the authors.

In sum, in the READY project of the collaborative research program 378, the following specific research results have been achieved for learning BNs in UASs:

- New BN learning algorithms have been developed, which are particularly well suited to the requirements of UASs.
- Existing and newly developed methods have been integrated into a general conceptualization of learning BNs for UASs.
- Important questions that have to be answered when applying ML techniques for UASs have been identified and treated by the development of new learning methods for BNs.
- The UMs of the READY prototypes have been empirically grounded by exploiting the data, that has been gathered in psychological experiments within READY, as input for the BN learning methods.

Notation: Bayesian Networks

Formally, a BN $B = (G, \theta)$ consists of two components. The first one is a directed acyclic graph G that represents the causal independencies that hold in the domain to be modeled by B. Nodes represent random variables and directed links between nodes are commonly interpreted as causal influences between these variables. We restrict our attention in this chapter to BNs in which all of the variables are discrete.

BNs are characterized by the following independence assumption: Given the states of its parents, a node is independent of all its nondescendents in the BN. The second component of a BN is a vector θ of conditional probability tables (CPTs) θ_i that represent the (uncertain) relationships between nodes and their parents. A node's CPT consists of conditional probabilities for each state of the node conditioned on its parents' state configuration. A BN represents a joint probability distribution $P(X_1, \ldots, X_n)$ over the states of its variables $\mathbf{X} = \{X_1, \ldots, X_n\}$. Exploiting the independence assumption of BNs, the joint probability distribution decomposes into a product of local conditional probabilities:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{pa}(X_i)) = \prod_{i=1}^{n} \theta_i. \tag{1}$$

The term $\mathbf{pa}(X_i)$ represents the set of all configurations of X_i's parents, while θ_i is the CPT belonging to node X_i. Therefore, $\theta = (\theta_1, \ldots, \theta_n)$. $\theta_{ijk} = P(X_i = x_{ij} \mid \mathbf{pa}(X_i) = pa_k(X_i)) = P(x_{ij} \mid pa_k(X_i))$stands for the entry corresponding to the jth state of X_i in θ_i when its parents take on their kth configuration $pa_k(X_i)$.

Structural Learning with Meta-Bayesian Networks

Generally, it is a hard problem to identify the single right structure by searching the large space of potential candidates on the basis of a given empirical dataset. This becomes worse as the size of the dataset decreases and in parallel the number of local optima increases. Therefore, model averaging is commonly applied [8], i.e., a set of good structures is maintained that are used together within the inference procedure weighted by their estimated quality. One drawback of that method – besides the question where the models originate from – is that it is not well suited with regard to the interpretability issue in the UM context. It is quite difficult to understand the reasoning process, i.e., which model contributed what to the final result. We developed a method that maintains an additional data structure that allows us to determine at any time the most promising BN UM that should be used for reasoning.

Standard BN structure learning methods based on a Bayesian scoring function yield a structure G accompanied with its (estimated) posteriori probability $P(G \mid \mathbf{D})$ conditioned on the available training data $\mathbf{D}$. This probability applies to G as a whole and states nothing about the likelihood of the presence of a particular link between two variables. Reference [6] present a method for estimating such

posteriori probabilities of links using MCMC methods. Meta-BNs, as described in the following, go a step further and represent a model of the joint probability distribution of the presence/absence of links that can be used for reasoning about the presence or absence of particular links conditioned on the presence/absence of other links.

Meta-BNs

We use meta-BNs for learning BN structures in the way introduced by [9].

A *meta-node* X^M_{vw} of a meta-BN $B^M = (G^M, \theta^M)$ represents a potential link between two nodes X_v and X_w of the application BN B. Each of these meta-nodes has three states, $x^M_{vw_{\leftrightarrow}}$, $x^M_{vw_{\rightarrow}}$, and $x^M_{vw_{\leftarrow}}$, which denote (i) the absence of a corresponding link, (ii) the presence of a link from node X_v to node X_w, and (iii) the presence of a link from X_w to node X_v, respectively.

Meta-links are links between meta-nodes of B^M that reflect direct dependencies between links in the application BN B. For example, the inclusion of one link may force another link to be removed from G. The meta-CPTs θ^M quantify these dependencies.

Figure 13 shows an example of a BN together with a potential meta-BN. The meta-BN models a direct relationship between the presence/absence of the links B $\rightarrow$ D and C $\rightarrow$ D, e.g., a 0.8 probability of C $\rightarrow$ D being present if B $\rightarrow$ D is absent.

In this way, a meta-BN B^M is able to model a probability distribution over the space of possible structures of B (on the basis of the potential links). It represents a meta-model of the relationships between different user properties. It can serve as an additional source to explore and analyze influencing aspects of the user behavior encoded in the UMs.

In the following, we will describe how to learn such a meta-BN B^M using available empirical data **D**.

Learning Meta-BNs

We present a method that extends Hofmann's framework—which can handle only very small domains with very few variables—in order to enable it to cope with

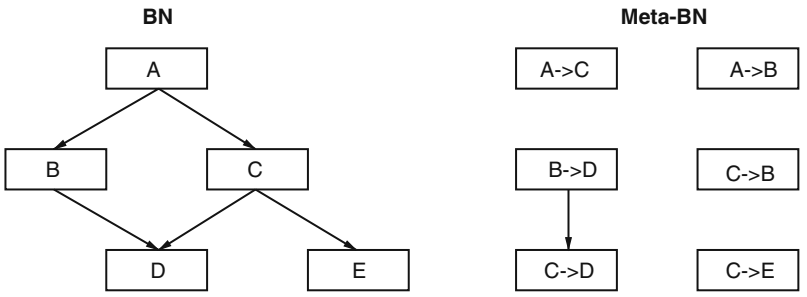


Fig. 13 Example of a meta-BN

domains involving more nodes and missing data, thereby making it an alternative for an application in scenarios beyond small artificial examples.

Addressing the issue of learning a meta-BN, we have to consider three partial learning tasks: (i) finding out which meta-nodes to include, i.e., determining which links to consider for possible inclusion in the application BN structure G; (ii) learning a DAG G^M to model the (in-)dependencies between the links of G; and (iii) learning the CPTs θ^M of B^M. We solve these problems in a two-step process; the last two issues are addressed together in the second step.

During the initial phase, we apply a standard BN structure learning method to learn a set $\mathbf{G}$ of m representative BN structures $\mathbf{G} = \{G_1, \ldots, G_m\}$ with posterior probabilities $P(G_i \mid \mathbf{D})$ as done in selective model averaging (see, e.g., [8]). Using this set of BNs, we determine the meta-nodes as follows: On the basis of the reasonable assumption that if a link plays a certain role in the user modeling process, it will be present in at least one of the learned structures of $\mathbf{G}$; we include a corresponding meta-node for each link that occurs in at least one of these m BNs. The likelihood that all relevant links occur in $\mathbf{G}$ can be increased by increasing the number m to learn BN structures.

After choosing the meta-nodes of the meta-BN B^M, it remains to learn the links of the meta-BN structure G^M and the corresponding CPTs θ^M. A solution to this problem is to use the set $\mathbf{G}$ of potential structures as a training dataset for the meta-learning task: Each structure G_i represents a set of statements regarding the absence and presence (with directions) of the potential links and therefore yields a training case for the meta-structure learning task. Each of these *meta-cases* is weighted by the (estimated) posterior probability $P(G_i \mid \mathbf{D})$ of the associated structure. Here again, standard BN structure learning methods are used. The underlying idea of such an approach is to compute the posterior probability $P(L \mid \mathbf{D})$ of a link $L \in \{x^M_{vw_{\rightarrow}}, x^M_{vw_{\leftarrow}}, x^M_{vw_{\leftrightarrow}}\}$. We extend Hofmann's learning method—exhaustive enumeration of all potential structures—by estimating the posterior probabilities by selective model averaging according to Eqs. (2) and (3) with $P(L \mid G_i, \mathbf{D}) = 1$ if L is present in G_i and $P(L \mid G_i, \mathbf{D}) = 0$ otherwise. This way, we enable meta-BNs to be used in more complex domains and/or in the presence of missing data.

$$P(L \mid \mathbf{D}) = \sum_{\mathbf{G}} P(L \mid G, \mathbf{D}) P(G \mid \mathbf{D}) \tag{2}$$

$$P(G \mid \mathbf{D}) \approx \frac{P(\mathbf{D} \mid G) P(G)}{\sum_{G_i \in \mathbf{G}} P(\mathbf{D} \mid G_i) P(G_i)} \tag{3}$$

Structural Adaptation with Meta-Bayesian Networks

Based on this meta-BN framework, we present a new structural adaptation algorithm for BNs that benefits from the compact encoding of the structural uncertainty in the meta-BNs.

```
STRUCTURAL ADAPTATION(D, s, k, m)
B^M ← learn_meta_BN(D, s, m)
B ← determine_BN(B^M)
while ¬exit do
    D^adapt ← ∅
    for i = 1 to k do
        case ← get_next_adaptation_case()
        D^adapt ← D^adapt ∪ case
        B ← adapt_CPTs(B, case)
    B^M ← adapt_CPTs_of_meta_BN(B^M, D^adapt, m)
    B ← determine_BN(B^M)
```

Fig. 14 Structural adaptation with meta-BNs

The basic idea of our approach is an application of standard CPT adaptation methods on the meta-level, i.e., the adaptation of the meta-BN's CPTs θ^M. Figure 14 shows the basic algorithm of our structural adaptation method. Free parameters that have to be specified manually are the global equivalent sample size (ESS) s to determine the rate at which the BN UM will be adapted, the number m of representative structures for the meta-learning procedure and the size k of the adaptation window.

The first step is to construct a meta-BN B^M on the basis of available empirical data $\mathbf{D}$. The information encoded in this meta-BN B^M is then used to choose the initial BN UM B. After a window of k new adaptation cases $\mathbf{D}^{adapt}$ (which are also used to adapt the current BN's CPTs according to standard CPT adaptation procedures), this new dataset is used as a basis for a single adaptation step of B^M's CPTs. The updated meta-information may or may not yield a structurally modified BN UM B.

Structural Adaptation Procedure

How can we use a meta-BN B^M to adapt the BN UM B to new observation about the user?

As was already indicated, we apply a standard CPT adaptation method to adapt the CPTs θ^M on the meta-level. To be able to do this, we have to transform the k adaptation cases $\mathbf{D}^{adapt}$ of the last adaptation window into a set of cases $\mathbf{D}^{M,adapt}$ that is appropriate for application with the meta-BN B^M.

To this end, we learn a set of m representative BNs on the basis of the k adaptation cases $\mathbf{D}^{adapt}$ in a manner analogous to the learning of the initial meta-BN. Together, these BNs are treated as a single particular observation in the domain for the meta-BN in a way we will describe in the following. On the basis of this observation, the meta-BN's CPTs are adapted according to the chosen CPT adaptation procedure.

For each meta-state $x^M_{vw_j}$, the posterior probability of the corresponding link $P(x^M_{vw_j} \mid \mathbf{D}^{adapt})$ is computed through the application of Eqs. (2) and (3) as described in Sect. 4. These $P(x^M_{vw_j} \mid \mathbf{D}^{adapt})$ are subsequently used as *likelihood evidences* for the meta-BN's meta-nodes and a meta-CPT adaptation step can be performed.

After this, the updated meta-BN's *most probable hypothesis* is computed. The result is a vector of meta-states and thus a vector of links that represent exactly one BN structure. The most probable structure, which obeys the constraint of representing an acyclic structure, is chosen to be the new adapted BN structure G' for the BN UM. It can be determined as follows: If the vector of links represents an acyclic structure, we are done. Otherwise, stochastic sampling with the meta-BN has to be performed to produce a set of acyclic structures, which is then used to estimate the most probable structure.

Finally, we need to choose appropriate CPTs θ' and ESSs $s'_{ik}$for the adapted BN structure G'. Most parts of the CPTs θ' remain unchanged after a structural adaptation step. Only those θ'_i that are related to a structural change have to be updated. These values can be computed as $P(X_i \mid pa_k^{old}(X_i))$ using the BN B^{old} before adaptation took place, by standard BN inference methods. In most realistic scenarios, the s'_{ik} needed for θ''s adaptation can be maintained in an additional data structure [16].

Particular Instantiation for the Evaluation

In the evaluation presented in this chapter, we made the following choices for the generic parts of the proposed structural adaptation framework: For structural learning, we used SEM [4] (with five random restarts in order to improve the quality of **G**, i.e., to avoid a set of very similar structures) with the Bayesian Information Criterion [22] as scoring function, we estimated the posterior probabilities as described in Sect. 4 using Eqs. (2) and (3) (using the $m = 60$ G_i with highest scores) and applied the AHUGIN procedure to adapt the CPTs. For the meta-structure learning task, we used SEM with a Bayesian scoring function that "punished" structures which included additional links by a factor of 0.9 (by specification of an appropriate prior probability distribution in Eq. (3) $P(G) \sim 0.9^{\#links}$). These particular choices represent a quite general instantiation of the framework. Dependent on the domain's demands, its performance can be improve by using especially well-suited learning algorithms or approximation methods.

References

1. Albrecht, D.W., Zukerman, I., Nicholson, A.E. Bayesian models for keyhole plan recognition in an adventure game. User Modeling and User-Adapted Interaction, 8:5–47 (1998).
2. Billsus, D., Pazzani, M.J. A hybrid user model for news story classification. In J. Kay (Ed.), UM99, User Modeling: Proceedings of the Seventh International Conference (pp. 99–108). Wien: Springer (1999).
3. Bohnenberger, T., Brandherm, B., Großmann-Hutter, B., Heckmann, D., Wittig, F. Empirically grounded decision-theoretic adaptation to situation-dependent resource limitations. Künstliche Intelligenz, 16(3):10–16 (2002).
4. Friedman, N. Learning belief networks in the presence of missing values and hidden variables. In: Proceedings of the 13th International Conference on Machine Learning (1997).

5. Friedman, N., Goldszmidt, M. Sequential update of Bayesian network structure. In D. Geiger, P.P. Shenoy (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference (pp. 165–174). San Francisco: Morgan Kaufmann (1997).
6. Friedman, N., Koller, D. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. Machine Learning, 50:95–126 (2002).
7. Heckerman, D. A tutorial on learning with Bayesian networks. In M.I. Jordan (Ed.), Learning in Graphical Models. Cambridge, MA: MIT Press (1998).
8. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. Bayesian model averaging: A tutorial. Statistical Science, 14:382–417 (1999).
9. Hofmann, R. Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen. Ph.D. thesis, Technische Universität München (2000).
10. Horvitz, E., Jacobs, A., Hovel, D. Attention-sensitive alerting. In K.B. Laskey, H. Prade (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference (pp. 305–313). San Francisco: Morgan Kaufmann (1999).
11. Horvitz, E., Koch, P., Kadie, C.M., Jacobs, A. Coordinate: Probabilistic forecasting of presence and availability. In A. Darwiche, N. Friedman (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (pp. 224–233). San Francisco: Morgan Kaufmann (2002).
12. Jameson, A., Großmann-Hutter, B., March, L., Rummer, R. Creating an empirical basis for adaptation decisions. In H. Lieberman (Ed.), IUI 2000: International Conference on Intelligent User Interfaces (pp. 149–156). New York, ACM (2000).
13. Jameson, A., Großmann-Hutter, B., March, L., Rummer, R., Bohnenberger, T., Wittig, F. When actions have consequences: Empirically based decision making for intelligent user interfaces. Knowledge-Based Systems, 14:75–92 (2001).
14. Jameson, A., Wittig, F. Leveraging data about users in general in the learning of individual user models. In B. Nebel (Ed.), Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (pp. 1185–1192). San Francisco, CA: Morgan Kaufmann (2001).
15. Lau, T., Horvitz, E. Patterns of search: Analyzing and modeling Web query dynamics. In J. Kay (Ed.), UM99, User Modeling: Proceedings of the Seventh International Conference (pp. 119–128). Wien: Springer (1999).
16. Moore, A., Lee, M.S. Cached sufficient statistics for efficient machine learning with large datasets. Journal of Artificial Intelligence Research, 8:67–91 (1998).
17. Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., Wittig, F. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In M. Bauer, P. Gmytrasiewicz, J. Vassileva (Eds.), UM2001, User Modeling: Proceedings of the Eighth International Conference. Berlin: Springer (2001).
18. Nicholson, A., Boneh, T., Wilkin, T., Stacey, K., Sonenberg, L., Steinle, V. A case study in knowledge discovery and elicitation in an intelligent tutoring application. In J. Breese, D. Koller (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (pp. 386–394). San Francisco: Morgan Kaufmann (2001).
19. Olesen, K.G., Lauritzen, S.L., Jensen, F.V. aHUGIN: A system creating adaptive causal probabilistic networks. In D. Dubois, M.P. Wellman, B. D'Ambrosio, P. Smets (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference (pp. 223–229). San Mateo: Morgan Kaufmann (1992).
20. Schäfer, R. Benutzermodellierung mit dynamischen Bayes'schen Netzen als Grundlage adaptiver Dialogsysteme. Ph.D. thesis, Lehrstuhl Wahlster, Fachrichtung Informatik, Universität des Saarlandes, Saarbrücken (1998).
21. Schäfer, R., Weyrath, T. Assessing temporally variable user properties with dynamic Bayesian networks. In A. Jameson, C. Paris, C. Tasso (Eds.), User Modeling: Proceedings of the Sixth International Conference, UM97 (pp. 377–388). Wien: Springer (1997).
22. Schwarz, G. Estimating the dimension of a model. Annals in Statistics, 6:461–464 (1978).
23. Webb, G., Pazzani, M.J., Billsus, D. Machine learning for user modeling. User Modeling and User-Adapted Interaction, 11:19–29 (2001).

24. Wittig, F. Maschinelles Lernen Bayes'scher Netze für benutzeradaptiver Systems. No. 267 in Dissertationen zur Künstliche Intelligenz. Akademische Verlagsgesellschaft Aka GmbH, Berlin (2003).
25. Wittig, F., Jameson, A. Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. In C. Boutilier, M. Goldszmidt (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (pp. 644–652). San Francisco: Morgan Kaufmann (2000).

Scope Underspecification with Tree Descriptions: Theory and Practice

Alexander Koller, Stefan Thater, and Manfred Pinkal

1 Introduction

Scope underspecification is the standard technique used in computational linguistics to deal efficiently with scope ambiguities, a certain type of semantic ambiguity. Its key idea is to not enumerate all semantic readings from a syntactic analysis during or after parsing as in more traditional approaches to semantic construction, but to derive a single, compact *underspecified representation (USR)* that *describes* the set of readings instead. The individual readings can be enumerated from an USR, but this process can be delayed until the readings are actually needed, say, for performing inference tasks. Furthermore, and more importantly, it is possible to perform certain types of inferences directly on the level of underspecification, i.e., *without* explicitly enumerating all readings described by an USR.

The establishment of scope underspecification as a standard technique is due in part to research in the project CHORUS, which took place in the context of the Collaborative Research Center (SFB) 378 at Saarland University from 1996 to 2007. Research in this project led to an improved understanding of the formal underpinnings of scope underspecification, the development of efficient algorithms for computing readings from underspecified descriptions, and practical methods and tools for underspecification in large corpora.

In this chapter, we outline some of these key results, highlighting especially two major lines of research. First, we review a series of solvers for *dominance constraints* and *dominance graphs*, two closely related underspecification formalisms developed in CHORUS. Our presentation of these increasingly fast solvers goes hand in hand with the development of increasingly deep insights into the structural properties exhibited by underspecified descriptions for natural language, and we end up with a clear view of the linguistically relevant fragment of dominance structures, together with a polynomial solver for this fragment – despite the fact that

A. Koller (✉)
MMCI and Department of Computational Linguistics and Phonetics, Saarland University, 66123 Saarbrücken, Germany
e-mail: koller@mmci.uni-saarland.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,* Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_15,

solving unrestricted dominance constraints is an NP-complete problem and therefore intractable.

Second, we present research which is concerned with semantic construction and the development of wide-coverage underspecification methods that can be applied on corpus data. We show how underspecified descriptions in the Minimal Recursion Semantics (MRS) formalism [9] can be translated into equivalent dominance graphs. This makes our fast solvers available to users of MRS, and makes large-scale grammars and annotated corpora producing MRS descriptions available as a resource for us. Although the translation of MRS to dominance graphs is only provably correct for the fragment of *hypernormally connected* underspecified descriptions, we show that virtually all (correctly modeled) MRS descriptions that appear in practice fall into this fragment. At the same time, investigating grammar rules that can produce MRS descriptions that are not hypernormally connected can be a powerful tool for debugging the semantic construction component of a grammar. We present an algorithm for reducing the logical redundancy of the readings represented by an USR and demonstrate its effectiveness and efficiency on MRS corpus data; and finally, we show some previously unpublished results on the annotation of scope ambiguities in a corpus.

The paper is structured as follows. We will first sketch the basic ideas behind scope underspecification and briefly review the basic definitions of dominance constraints and dominance graphs in Sect. 2. In Sect. 3, we then review a series of four solvers for dominance constraints and graphs, ranging from a purely logic-based saturation algorithm over a solver based on constraint programming to efficient solvers based on graph algorithms. In Sect. 4, we show how our semantic techniques can be linked to deep grammatical processing and wide-coverage grammars, and also review an algorithm for eliminating redundant readings from underspecified descriptions. Section 5 then presents some previously unpublished results on corpus-based studies we carried out in order to obtain a statistical model of scope. We conclude in Sect. 6.

2 Dominance-Based Scope Underspecification

The fundamental problem that scope underspecification approaches set out to solve is to manage the readings of sentences with scope ambiguities efficiently. Scope ambiguity is a specific type of semantic ambiguity. For instance, sentence (1) is ambiguous between reading (3) (in which all students read the same book) and reading (2) (in which they may be reading different ones):

(1) Every student reads a book.
(2) $\forall x.student(x) \rightarrow (\exists y.book(y) \wedge read(x, y))$.
(3) $\exists y.book(y) \wedge (\forall x.student(x) \rightarrow read(x, y))$.

The number of readings can grow exponentially in the number of quantifiers or other scope-bearing operators (e.g., negations or modal operators) occurring in the sentence, so simple approaches to semantic interpretation that first enumerate all

readings and then select the intended one(s) cannot be efficient in general. The problem is illustrated by sentence (4), which is predicted to have about $2 \times 5!$ (=14400) readings if we assume that quantifiers must take scope within their own clause.

(4) A politician can fool most voters on most issues most of the time, but no politician can fool every voter on every single issue all of the time [39].

Large-scale grammars that can do semantics construction, such as the English Resource Grammar [8], routinely predict millions of scope readings for a sentence. For instance, one sentence from the Rondane corpus (see Sect. 4) was found to have 2.4 trillion readings according to a certain version of the grammar. While not all these readings represent genuine meaning differences, the problem of managing so many readings efficiently remains.

Underspecification is a standard technique to address this problem [40, 6, 9]. Its key idea is to not enumerate all semantic readings from a syntactic analysis during or after parsing, but to derive a single, compact *underspecified description,* or *underspecified representation (USR),* instead. Current algorithms (see Sect. 3) support the efficient enumeration of readings from underspecified descriptions. However, the true promise of using underspecified descriptions is that they may be a useful platform for reducing the set of described readings to those that could actually have been meant in the given context, *without* enumerating these readings explicitly (see Sect. 4). Finally, underspecification simplifies the specification of the syntax–semantics interface, and all large-scale grammars with a hand-crafted syntax-semantics interface that we are aware of use some form of underspecification (e.g., [8, 4, 11]).

When the CHORUS project started in 1996, several underspecification formalisms, such as UDRT [40], Hole Semantics [6], and MRS [9] already existed, and MRS-based wide-coverage grammars were under development. However, these formalisms emphasized the conceptual exploration of underspecification, rather than rigorous formalization and efficient algorithms.

Against this background, the CHORUS project developed the underspecification formalisms of context unification [34], the Constraint Language for Lambda Structures (CLLS) [15], dominance constraints, and dominance graphs [1]. Each of these formalisms provides well-defined mechanisms for describing sets of trees and exploits the fact that semantic representations like (2) and (3) can be represented as trees (see Fig. 1); the formalisms are closely related to each other, and we will point out the relationships where appropriate. We will now briefly review dominance constraints and dominance graphs.

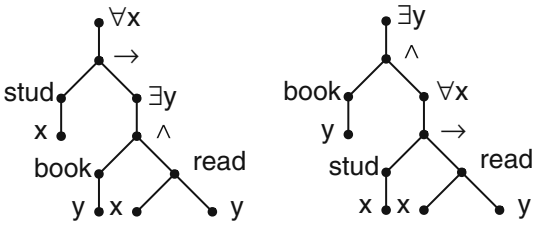


Fig. 1 Trees for the readings (2) and (3)

2.1 Dominance Constraints

Dominance constraints are first-order formulas that talk about the parent–child relation and the dominance relation (its reflexive, transitive closure) between nodes in a tree. Dominance-based tree descriptions were first used in automata theory in the 1960s [43], rediscovered in computational linguistics in the 1980s [30], and studied in the early 1990s from a logical point of view [3]. They have found numerous applications in computational linguistics – e.g., for grammar formalisms [41], natural language semantics [15], and discourse [20] – and other areas, such as XML database theory [21].

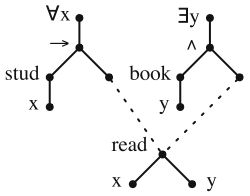
For underspecified semantics, the basic idea is to consider the readings of ambiguous natural language expressions as trees and to compactly describe these trees by specifying what they have in common. An example of a dominance constraint is shown in Fig. 2 on the right. It consists of a conjunction of *labeling literals* (e.g., $X_1 : \forall x(X_2)$) and *dominance literals* of the form $X \lhd^* Y$. The graph to its left shows a graphical representation of the constraint, where the trees defined by the solid edges stand for labeling literals, and the dotted lines stand for dominance literals. The constraint can be seen as an underspecified description of the two trees in Fig. 1: Labeling literals specify the "semantic material" the two trees consist of, and the two dominance literals require that the nuclear scope ("read") must be within the scope of the two quantifiers. Because both quantifiers outscope the same nuclear scope, one of them must outscope the other; therefore, the two trees in Fig. 1 are the only two arrangements of the graph.

Notice that from the perspective of the dominance constraint, symbols like "$\forall x$" and "$\rightarrow$" that occur to the right of the colon in a labeling literal are simply uninterpreted node labels in the tree. The fact that these symbols have a special meaning when reading the tree as a representation of a predicate logic formula is irrelevant at this level, and in particular the lowercase x in $\forall x$ is not a variable in the same sense as the uppercase X_1 and X_2, but simply part of the label.

More formally, the syntax of dominance constraints is defined as follows, where f is an n-ary function symbol from some given signature, and X, Y, X_i (for $1 \leq i \leq n$) are variables.

$$\varphi := X{:}f(X_1, \ldots, X_n) \mid X \lhd^* Y \mid X \neq Y \mid \varphi \wedge \varphi'$$

Dominance constraints are interpreted over (model structures representing) finite ordered constructor trees, i.e., trees in which the arities of the node labels determine



$X_1 : \forall x(X_2) \wedge X_2 : \rightarrow(X_3, X_4) \wedge$
$X_5 : stud(X_6) \wedge X_6 : x \wedge$
$X_{10} : \exists y(X_{11}) \wedge X_{11} : \wedge(X_{12}, X_{13}) \wedge$
$X_{12} : book(X_{14}) \wedge X_{14} : y$
$X_4 \lhd^* X_7 \wedge X_{13} \lhd^* X_7$
$X_7 : read(X_8, X_9) \wedge$
$X_8 : x \wedge X_9 : y$

Fig. 2 A dominance constraint (*right*) and its graphical representation (*left*); the solutions of the constraint are the two trees in Fig. 1

the number of outgoing edges; alternatively, such trees can be seen as ground terms over a ranked signature. The variables denote nodes in such a tree. A *labeling atom* $X : f(X_1, \ldots, X_n)$ expresses that the node denoted by variable X has label f, and its children are the nodes denoted by $X_1, \ldots, X_n$; the *dominance atom* $X \lhd^* Y$ says that there is a path from the node denoted by X to the node denoted by Y; and the *inequality atom* $X \neq Y$ means that X and Y denote different nodes. A tree *satisfies* a dominance constraint iff there is an assignment of nodes to variables such that each atom is satisfied. In this case, we say that the pair of the tree and the assignment is a *solution* of the constraint. For instance, the two tree in Fig. 3 satisfy the constraint on the right of Fig. 2.

Dominance constraints can be extended to more powerful languages, such as the Constraint Language for Lambda Structures (CLLS; [15]), which adds parallelism and binding constraints. Parallelism constraints can be used to model the interaction of scope and ellipsis, and binding constraints account for variable binding without using variable names to avoid unwanted "capturing" of variables, which is problematic in particular if parallelism constraints are used. Dominance and parallelism constraints together are equivalent to context unification, a formalism we used in CHORUS before developing CLLS [33]. Furthermore, our use of dominance constraints was one factor in the development of Extensible Dependency Grammar (XDG; Dependency Grammar by Debusmann and Kuhlmann, this volume).

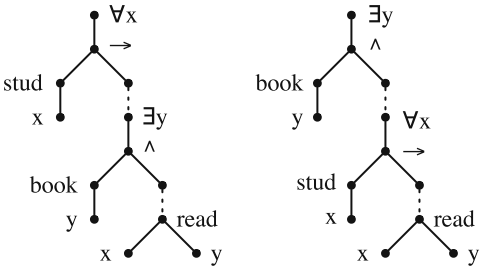


Fig. 3 Solved forms of the dominance graph in Fig. 2

2.2 Dominance Graphs

An alternative way of looking at the graph in Fig. 2 is to read it directly as a *dominance graph* [1]. A dominance graph is a directed graph with two kinds of edges – tree edges (drawn solid) and dominance edges (drawn dotted) – such that the graph is a set of trees if all dominance edges are deleted. These trees are the *fragments* of the graph, and we can define the notions "root" and "leaf" relative to these fragments. We usually consider *labeled* dominance graphs, in which each non-leaf is assigned a label; leaves can be unlabeled, in which case they are called *holes.*

Intuitively, a tree is a *solution* of a dominance graph G iff the graph can be embedded into the tree. More formally, a solution of G consists of a pair (t, α), where t is a tree and α a function mapping the nodes in G to nodes in t, such that

no two labeled nodes are mapped to the same node in t, all labels and tree edges are preserved, and all dominance edges are realized as reachability in t. This is clearly the case for the trees in Fig. 1 and the graph in Fig. 2.

In general, it is possible to translate every dominance graph into a dominance constraint with the same solutions (see Fig. 2). Conversely, all *overlap-free* dominance constraints can be translated into equivalent dominance graphs. A constraint is called overlap-free if for any two labeling atoms $X{:}f(\ldots)$ and $Y{:}g(\ldots)$, where X and Y are different variables but f and g not necessarily different labels, it also contains an inequality atom $X \neq Y$.

The primary advantage of dominance graphs over dominance constraints is that the graph-perspective on scope underspecification leads in a natural way to restrictions that allow us to process underspecified descriptions efficiently: For instance, [1] identify the fragment of *normal* dominance graphs and show that the solvability problem – does the graph have a solution? – can be decided in polynomial time. Recently, more general classes of dominance graphs which can still be solved efficiently have been identified [5, 29].

2.3 Configurations and Solved Forms

The solution of a dominance graph or constraint may contain nodes that were not mentioned in the graph (i.e., they are not denoted by any node or variable), which means that in general, every solvable dominance graph has an infinite number of solutions. This is a desirable feature in some cases, as it allows us, for instance, to monotonically add more semantic material in order to model reinterpretation of metonymic expressions or ellipsis reconstruction [14]. But there cannot possibly be an algorithm for enumerating all solutions of a graph, and for cases where the graph only represents a scope ambiguity, we would still like to be able to construct semantic representations that only consists of the semantic material that was mentioned in the sentence.

There are two ways to work around this discrepancy. One is to look not at all solutions of a graph, but only at its *configurations*. A solution (t, α) of a graph G is called a configuration of G iff every node of t is the α-image of a non-hole in G. In other words, every node in the tree is the image of a labeled node in the graph, i.e., all node label choices in the tree are forced by the embedding. The two trees in Fig. 1 are the only two configurations of the graph in Fig. 2.

Alternatively, we can choose to consider not all solutions of a dominance graph, but its *solved forms*. A dominance graph is called a solved form iff it is a forest, i.e., it has no directed cycles and no node has two parents. A graph G is called a *solved form of* some other graph G' iff G is in solved form, the two graphs have the same nodes, tree edges, and node labels, and for each dominance edge (u, v) in G', there is a directed path from u to v in G. Every dominance graph has only a finite number of solved forms, and all dominance graphs in solved form are solvable; the solved forms of a dominance graph partition the (infinitely many) solutions of the graph into a finite set of classes, such that the solutions in each class only differ in details

the graph does not have anything to say about in the first place. In particular, the graph in Fig. 2 has exactly two solved forms, which are shown in Fig. 3.

The choice between considering configurations and solved forms depends on the perspective on underspecification. If we only care about the semantic representations that are built from the explicitly mentioned semantic material, we want the configurations of the dominance graph. However, it is computationally more efficient to compute solved forms: Although it can be decided in polynomial time whether a dominance graph has a solved form [29], the question of whether a dominance graph has a configuration is NP-complete [1]. This is because there are dominance graphs which have solutions but no configurations; an example is shown in Fig. 4. However, it can be shown that for dominance graphs which are *hypernormally connected* (hnc) [24] the difference between solvability and configurability disappears: Every solved form of a normal, hypernormally connected graph has a (unique) configuration. A normal dominance graph is called hnc iff each pair of nodes is connected by a *simple hypernormal path*, i.e., a simple undirected path that never uses two dominance edges that are incident to the same hole. We will see below (Sect. 4) that there is strong evidence that all dominance graphs needed to model scope underspecification are hnc, so we can use polynomial solvers to compute the linguistically relevant configurations.

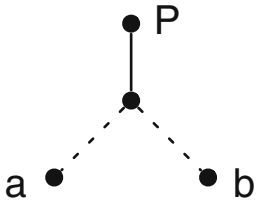


Fig. 4 A solvable dominance graph without configurations

3 Solving Dominance Constraints and Graphs

Dominance graphs and constraints provide clean formalisms for writing down underspecified semantic representations. But how can we retrieve the set of solved forms of an underspecified description efficiently? This is the *enumeration problem*; algorithms for the enumeration problem are called *solvers*. In general, a dominance constraint or graph may have an exponential number of solved forms, so even the best solvers must take exponential runtime. In order to get a more fine-grained analysis of a solver's efficiency, one can also consider the *satisfiability problem* of dominance graphs and dominance constraints, for which one only needs to decide whether the graph has any solved forms.

One of the key contributions of the CHORUS project was the successive improvement of solvers by identifying practically useful fragments of dominance constraints for which faster solvers could be developed. The progress was dramatic, both theoretically – the satisfiability problem for unrestricted dominance constraints is NP-complete, whereas the satisfiability problem for normal dominance graphs can be solved in linear time – and in practice. For instance, Fig. 5a shows the runtimes of

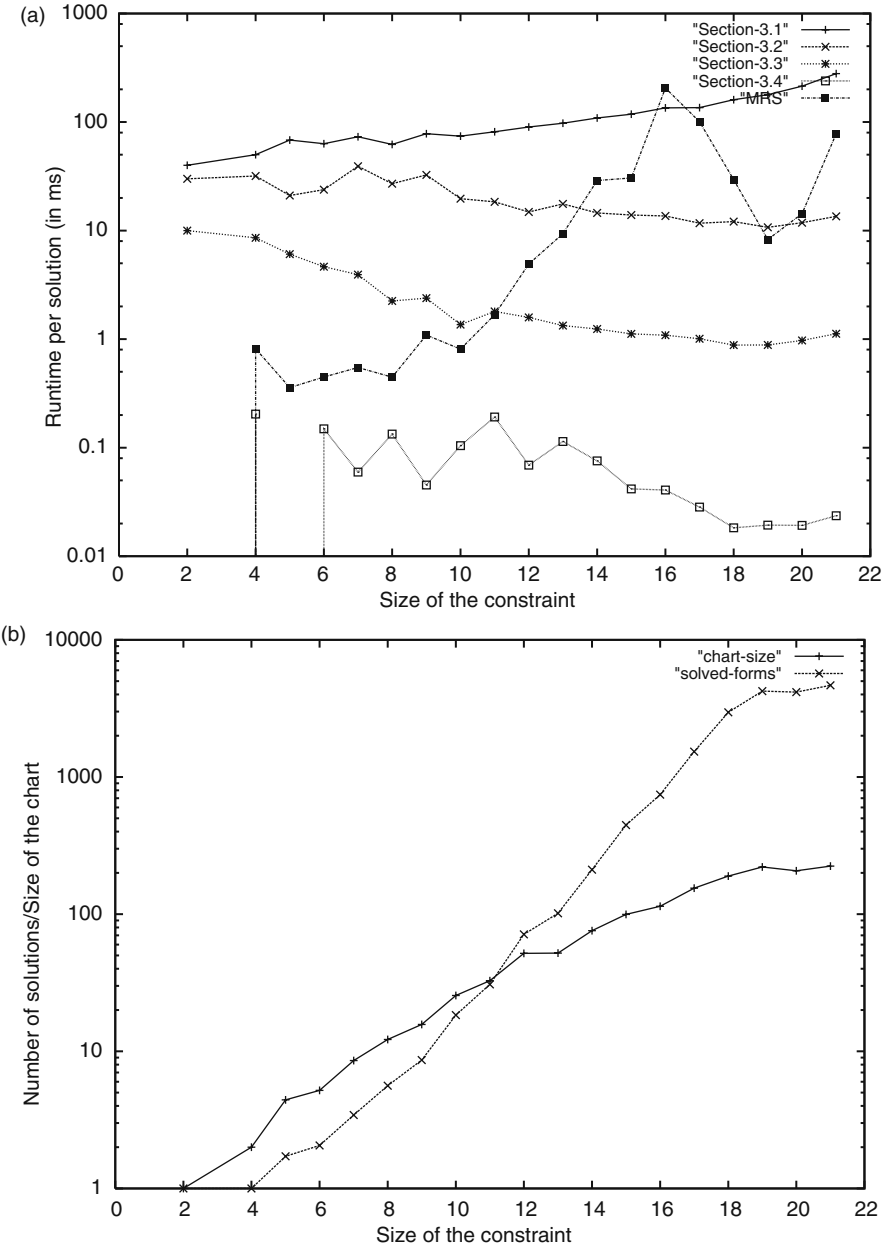


Fig. 5 (**a**) Runtimes per solved form for each of the solvers presented here, averaged over all dominance graphs of a given size in the Rondane treebank. (**b**) Comparison of the average chart sizes and average numbers of readings for each graph size

different underspecification solvers on MRS descriptions from the Rondane treebank (see Sect. 4). Each data point is the runtime of a solver divided by the number of solved forms it computes, averaged over all USRs of a certain size (number of fragments) in the treebank. As the runtime graph shows, each new solver is an order of magnitude faster than its predecessor. The standard MRS solver (from the LKB system [8]) is very fast for small descriptions, but its average runtime per reading grows much faster than is the case for the dominance-based solvers.

The high-level picture of the evolution of dominance constraint and graph solvers can be seen as follows:

- The first solver for CLLS was a saturation algorithm by Niehren, Erk, and Koller which operated directly on the constraints as logical formulas [25, 16]. As a solver for (unrestricted) dominance constraints, it solves an NP-complete problem and therefore has worst-case exponential runtime.
- Over the next years, Duchier and Niehren developed a solver for dominance constraints that worked by translating them into finite set constraints and then applying constraint programming techniques [13]. Although this solver still applies to general dominance constraints, it turned out that its search process never failed, and therefore ran in de facto polynomial time, on dominance constraints as used in underspecification. The behavior of the solver indicated that there was a – still unknown – polynomial time fragment of dominance constraints which subsumes all the underspecified representations of NL sentences the solver was tested with.
- One of these fragments was identified by Mehlhorn and Thiel in 2000, who presented a solver for *normal dominance graphs*, i.e., dominance graphs in which all dominance edges go from holes to roots [1]; these are equivalent to normal dominance constraints. An improved version of their solver was later shown to decide satisfiability of normal graphs in linear time.
- In 2004, Bodirsky, Niehren, and Miele presented a fundamentally different graph solver which was applicable to *weakly normal* dominance graphs, i.e., dominance graphs in which all dominance edges go into roots [5]. Their solver decides satisfiability in quadratic time, but is often more efficient than the Thiel solver in practice.
- Koller and Thater improved the efficiency of the Bodirsky solver by tabulating intermediate results in a chart data structure in 2005 [27] and generalized it to unrestricted dominance graphs in 2007 [29]. Their solver decides satisfiability of arbitrary dominance graphs in cubic time, but still solves satisfiability of weakly normal graphs in quadratic time.

We will now sketch the saturation solver, the set constraint solver, the Bodirsky solver, and the chart-based version of the Bodirsky solver, in turn.

3.1 A Saturation Algorithm

The first dominance constraint solver [25, 13] is an algorithm that operates directly on the constraint as a logical formula. It is a *saturation algorithm*, which successively enriches the constraint using *saturation rules*. The algorithm terminates if it

either derives a contradiction (marked by the special atom false), or if no rule can contribute any new atoms. In the first case, it claims that the constraint is unsatisfiable; in the second case, it reports the end result of the computation as a *solved form* and claims that it is satisfiable.

The saturation rules in the solver try to match their preconditions to the constraint, and if they do match, add their conclusions to the constraint. For example, the following rules express that dominance is a transitive relation, and that trees have no cycles:

$$\begin{aligned} X \lhd^* Y \wedge Y \lhd^* Z &\rightarrow X \lhd^* Z \\ X{:}f(\ldots, Y, \ldots) \wedge Y \lhd^* X &\rightarrow \text{ false} \end{aligned}$$

Some rules have disjunctive right-hand sides; if they are applicable, they perform a case distinction and add one of the disjuncts. One example is the *Choice Rule*, which looks as follows:

$$X \lhd^* Z \wedge Y \lhd^* Z \rightarrow X \lhd^* Y \vee Y \lhd^* X$$

This rule checks for the presence of two variables X and Y that are known to both dominate the same variable Z. Because models must be trees, this means that X and Y must dominate each other in some order; but we cannot know yet whether it X dominates Y or vice verse. Hence the solver tries both choices. This makes it possible to derive multiple solved forms (one for each reading of the sentence), such as the two different trees in Fig. 1.

It can be shown that a dominance constraint is satisfiable iff it is not possible to derive false from it using the rules in the algorithm. In addition, every model of the original constraint satisfies exactly one solved form. So the saturation algorithm can indeed be used to solve dominance constraints. However, even checking satisfiability takes nondeterministic polynomial time. Because all choices in the distribution rule applications have to be checked, a deterministic program will take exponential time to check satisfiability in the worst case. Indeed, satisfiability of dominance constraints is an NP-complete problem [25], and hence it is likely that any solver for dominance constraints will take exponential worst-case runtime.

3.2 Reduction to Set Constraints

In reaction to this NP-completeness result, Duchier and Niehren [13] applied techniques from *constraint programming* to the problem in order to get a more efficient solver. Constraint programming [2] is a standard approach to solve NP-complete combinatorial problems. In this paradigm, a problem is modeled as a formula in a logical constraint language. The program searches for values for the variables in the formula that satisfy the formula. In order to reduce the size of the search space, it performs cheap deterministic inferences that exclude some values of the variables

(propagation), and only after propagation can supply no further information it performs a non-deterministic case distinction *(distribution)*.

Duchier and Niehren solved dominance constraints by encoding them as *finite set constraints*. Finite set constraints [32] are formulas that talk about relations between (terms that denote) finite sets of integers, such as inclusion $X \subseteq Y$ or equality $X = Y$. Duchier and Niehren's implementation of their solver used the Mozart/Oz programming system [37], which comes with an efficient solver for set constraints.

The basic idea underlying the reduction is that a tree can be represented by specifying for each node v of this tree which nodes are dominated by v, which ones dominate v, which ones are equal to v (i.e., just v itself), and which ones are "disjoint" from v (Fig. 6). These four node sets are a partition of the nodes in the tree.

Now the solver introduces for each variable X in a dominance constraint φ four variables Eq_X, Up_X, $Down_X$, and $Side_X$ for the sets of node variables that denote nodes in the respective region of the tree, relative to X. The atoms in φ are translated into constraints on these variables. For instance, a dominance atom $X \lhd^* Y$ is translated into

$$Up_X \subseteq Up_Y \wedge Down_Y \subseteq Down_X \wedge Side_X \subseteq Side_Y$$

This constraint encodes that all variables whose denotation dominates the denotation of X (Up_X) must also dominate the denotation of Y (Up_Y), and the analogous statements for the dominated and disjoint variables.

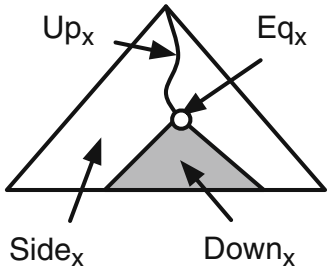


Fig. 6 The four node sets

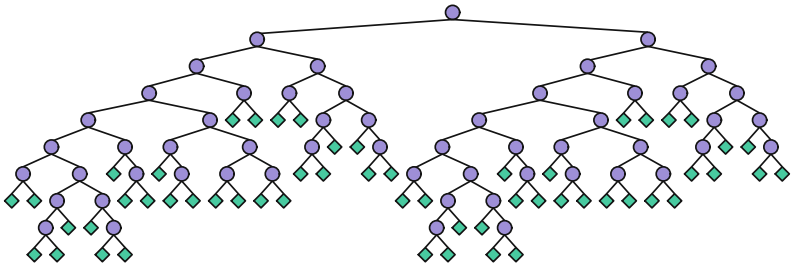


Fig. 7 Search tree for sentence 42 from the Rondane treebank

In addition, the constraint program contains various redundant constraints that improve propagation. Now the search for solutions consists in finding satisfying assignments to the set variables. The result is a search tree as shown in Fig. 7: The blue circles represent case distinctions, whereas each green diamond represents a solution of the set constraint (and therefore, a solved form of the dominance constraint). Interestingly, all leaves of the search tree in Fig. 7 are solution nodes; the search never runs into inconsistent constraints. This seems to happen systematically when solving any constraints used to model scope underspecification.

3.3 A Graph-Based Solver

This behavior of the set-constraint solver is extremely surprising: The key characteristic of an NP-complete problem is that there is no a priori bound on the number of failed nodes in the search tree. The fact that the solver never runs into failure is a strong indication that there is a fragment of dominance constraints that contains all constraints that are used to model scope underspecification, and that the solver automatically exploits this fragment. This insight led to the development of dominance graphs, which capture the overlap-free fragment of dominance constraints (which is sufficient for underspecification) and can be solved in worst-case polynomial time.

The most recent graph solver, developed by Bodirsky et al. [5], is a recursive procedure that successively splits a weakly normal dominance graph into smaller parts, solves them recursively, and combines them into complete solved forms. In each step, the algorithm identifies the *free fragments* of the dominance (sub-)graph. A fragment is free if it has no incoming dominance edges, and all of its holes are in different biconnected components of the undirected version of the dominance graph; this means that each undirected path that connects any two holes of a fragment goes through the root of this fragment. It then iterates over all free fragments F, removes each of them from the current subgraph G' in turn, and calls itself recursively on the weakly connected components of $G' - F$. It can be shown that if the subgraph that gets passed to any recursive call of the solver is unsolvable, then the entire original graph was unsolvable as well. This means that we can prove that the search for solved forms never fails, and together with the fact that each recursive call spends only linear time on computing the free fragments, this means that the entire algorithm can enumerate the N minimal solved forms of a dominance graph with n fragments in time $O(n^2 N)$.

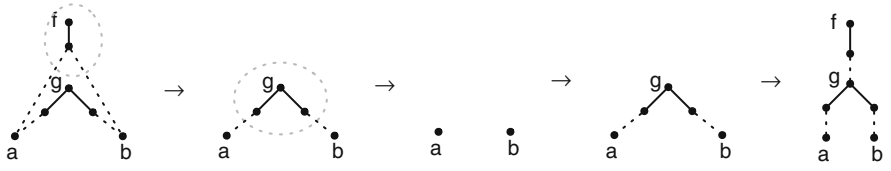


Fig. 8 An example computation of the graph solver

An example computation of the graph solver is shown in Fig. 8. The input graph is shown on the left. It contains exactly one free fragment F; this is the fragment whose root is labeled with f. (The single-node fragments both have incoming dominance edges, and the two holes of the fragment with label g are in the same biconnected component.) So the algorithm removes F from the graph and recursively solves the restricted graph. This restricted graph has again exactly one free fragment (the one labeled with g), so the algorithm removes this fragment and calls itself again recursively on the resulting two weakly connected components of the restricted subgraph. These two components both consist of single nodes, so we are finished. Finally the algorithm builds a solved form for the whole graph by first plugging the two single nodes under the two holes of the g-fragment, which is then plugged into the single hole of the f-fragment. The final result is shown on the right. By contrast, the graph in Fig. 9 has no solved forms. The solver will recognize this immediately, because none of the fragments are free (they either have incoming dominance edges, or their holes are biconnected).

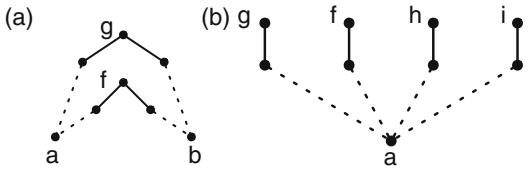
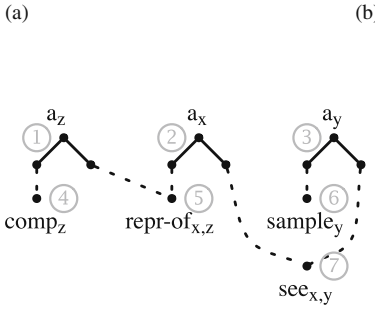


Fig. 9 **(a)** An unsolvable dominance graph; **(b)** A worst-case graph for the chart solver

3.4 The Chart Solver

Although the graph solvers were great leaps forward in solving dominance graphs efficiently, they have one weakness: If during the recursive computation they are called on the same subgraph multiple times, they repeat the same computation each time. In solving, for instance, the graph shown in Fig. 10, the Bodirsky solver will solve the subgraph consisting of the fragments $\{3, 6, 7\}$ twice, because it can pick the fragments 1 and 2 in either order. This is no big deal in this particular example because the subgraph is in solved form; but if we imagine that it is a larger subgraph that has many solved forms itself, the solver could waste a lot of time recomputing these solved forms a second time.

Koller and Thater [27] exploit this insight in an algorithm that uses dynamic programming techniques to store intermediate results of the Bodirsky solver in a data structure called a *dominance chart* (in analogy to charts as used in chart parsing). This data structure maps subgraphs of a dominance graph to a set of *splits* for this subgraph. Splits encode the splittings of the graph into weakly connected components that take place when a free fragment is removed. If F is a free fragment of the subgraph G and the weakly connected components of $G - F$ are $G_1, \ldots, G_n$, then the split induced by F in G is $\langle F, G_1 \mapsto u_1, \ldots, G_n \mapsto u_n \rangle$, where u_i is the lowest node in F from which a dominance edge points into G_i. In a normal



(b)

$$
\begin{aligned}
\{1,2,3,4,5,6,7\} :& \langle 1, h_1 \mapsto \{4\}, h_2 \mapsto \{2,3,5,6,7\}\rangle \\
& \langle 2, h_3 \mapsto \{1,4,5\}, h_4 \mapsto \{3,6,7\}\rangle \\
& \langle 3, h_5 \mapsto \{5\}, h_6 \mapsto \{1,2,4,5,7\}\rangle \\
\{2,3,5,6,7\} :& \langle 2, h_3 \mapsto \{5\}, h_4 \mapsto \{3,6,7\}\rangle \\
& \langle 3, h_5 \mapsto \{6\}, h_6 \mapsto \{2,5,7\}\rangle \\
\{3,6,7\} :& \langle 3, h_5 \mapsto \{6\}, h_6 \mapsto \{7\}\rangle \\
\{2,5,7\} :& \langle 2, h_3 \mapsto \{5\}, h_4 \mapsto \{7\}\rangle \\
\{1,4,5\} :& \langle 1, h_1 \mapsto \{4\}, h_2 \mapsto \{5\}\rangle \\
\{1,2,4,5,7\} :& \langle 1, h_1 \mapsto \{4\}, h_2 \mapsto \{2,5,7\}\rangle \\
& \langle 2, h_3 \mapsto \{1,4,5\}, h_4 \mapsto \{7\}\rangle
\end{aligned}
$$

Fig. 10 (**a**) A dominance graph with five equivalent readings and (**b**) its chart

dominance graph, all u_i are holes. The algorithm can now be more efficient than the Bodirsky solver by checking at the beginning of each recursive call if it has visited the subgraph before. If yes, it returns immediately; if not, it proceeds with the call as before, except that instead of returning a set of solved forms, it records the split for each free fragment in the chart.

Let us look at an example to make this clearer. Fig. 10b displays the chart that the algorithm computes for the graph in Fig. 10a. In the entire graph G (represented by the set $\{1, \ldots, 7\}$ of fragments), the fragments 1, 2, and 3 are free. As a consequence, the chart contains a split for each of these three fragments. If we remove fragment 1 from G, we end up with a weakly connected graph G_1 containing the fragments $\{2, 3, 5, 6, 7\}$. There is a dominance edge from the second hole h_1 of 1 into G_1, so once we have a solved form of G_1, we will have to plug it into h_1 to get a solved form of G; therefore G_1 is assigned to h_1 in the split. On the other hand, if we remove fragment 2 from G, G is split into two weakly connected components $\{1, 4, 5\}$ and $\{3, 6, 7\}$, whose solved forms must be plugged into the holes h_3 and h_4 of the fragment 2, respectively. Notice that the subgraph $\{3, 6, 7\}$ is referenced by two different splits, but its own splits are represented only once in the chart. In other words, the chart solver avoids performing the same computation (solving $\{3, 6, 7\}$) multiple times.

One extremely interesting recent connection is that dominance charts of hypernormally connected graphs can be seen as specific *regular tree grammars* [26], a standard device used in theoretical computer science for describing sets of trees [7]. Arguably, regular tree grammars could be seen as an underspecification formalism in their own right: In the worst case, the chart of a graph with n fragments can contain $O(2^n)$ splits (see Fig. 9); but this is still much less than the $O(n!)$ solved forms the graph may have, and in practice a set of trillions of readings can be represented by a chart with thousands of splits (see also Fig. 5b). Because regular tree grammars are a more explicit representation of the trees than dominance graphs and they potentially support algorithms that would be hard to realize on graphs (see also Sect. 4.3), exploring them as a tool for underspecification is an interesting avenue for future research.

4 Practical Scope Underspecification

The efficient solvers presented in the previous section are one crucial ingredient for making scope underspecification useful for practical applications. However, using underspecification in practice requires further components, including a syntax-semantics interface for wide-coverage grammars. Furthermore, scope underspecification is really only useful if USRs can be used for disambiguating inferences that eliminate readings that were not meant in the context. In this section, we report on some research along these lines.

First, we show how to obtain dominance graphs from sentences using wide-coverage grammars. Rather than extending our own hand-written construction rules for smaller grammars [15, 12], we showed how USRs in the MRS formalism [9] can be translated into dominance graphs (see Sect. 4.1). This immediately makes the wide-coverage HPSG grammars that compute MRS descriptions (such as the English Resource Grammar, or ERG [8]) and corpora that are annotated according to these grammars (such as the Rondane and Redwoods corpora [36]) available for us. Conversely, our solvers can now be used to work with HPSG grammars and allow us to perform tasks very efficiently that would have been impossible previously: For instance, using the chart solver from Sect. 3.4, we can compute the number of scope readings for all sentences in the Rondane treebank in about half a minute.

However, the translation is only correct for MRSs that translate into *hypernormally connected* dominance graphs. In Sect. 4.2, we report on experiments with the HPSG treebanks that support what we call the *Net Hypothesis*: That all USRs that are used in practice are hypernormally connected. In earlier versions of the treebanks, this claim was only true for 70–80% of all USRs, but by inspecting the grammar rules that contributed to the derivation of non-nets, we could identify rules with faulty semantic construction components. By fixing these rules, the percentage of nets has grown to about 97% of the treebanks in the most recent version. In other words, the Net Hypothesis is correct enough to be useful for grammar debugging.

To round off this section, we briefly sketch an algorithm for *redundancy elimination*, which strengthens an USR in order to remove readings that are logically equivalent to other readings, in Sect. 4.3. Such an algorithm is practically useful because an application will not care about apparent ambiguities which only represent syntactic variations on semantically equivalent readings. Our implementation reduces the median number of readings in the Rondane treebank from 56 to 4, and does this in negligible runtime. It crucially relies on the chart solver from Sect. 3.4 as well as on our research on hypernormally connected dominance graph, and thus brings the two strands of research presented here together.

4.1 Minimal Recursion Semantics as Dominance Constraints

Minimal Recursion Semantics (MRS) [9] is a formalism for modeling scope that is very similar to dominance constraints and graphs. The following example shows an MRS description for sentence (1).

(5) $h_1{:}every_x(h_2, h_3), h_4{:}student(x), h_5{:}some_y(h_6, h_7), h_8{:}book(y), h_9{:}read(x, y), h_2 =_q h_4, h_6 =_q h_8$

This MRS description (or simply *MRS* for short) consists of *elementary predications (EPs)* $h : F$ which specify the "semantic material" common to both readings, and *handle constraints* $h =_q h'$ which restrict the way EPs can be combined into a complete reading. In an EP $h : F$, the handle h is said to be the *label* of the EP; handles on the right hand side of a colon are also referred to as *argument handles*. The two EPs labeled by h_1 and h_5 stand for object language quantifiers[1] binding variables x and y, respectively. The readings of an MRS are those formulas that can be obtained from the MRS by "plugging" (instantiating) argument handles with labels, so that every argument handle is plugged by a label, and all but one label get plugged into some argument handle. The pluggings must be consistent with the handle constraints – if a handle constraint $h =_q h'$ occurs in an MRS, then the formula labeled by h' must be a subformula of the one labeled by h[2] – and all occurrences of bound object language variables must be in the scope of the quantifier that binds the variable.

If we compare (5) and the corresponding dominance constraint in Fig. 2, we can observe that the two underspecified descriptions are almost identical if we ignore minor details in the way quantifiers are represented in the two formalisms, so it is straightforward to define a translation of MRS into dominance constraints: Elementary predications correspond to labeling literals, handle constraints to dominance literals, and variable binding induces further dominance atoms from the quantifier to the variables with the same name. The resulting dominance graph for the example is shown on the left of Fig. 11.[3]

This translation is indeed correct in the sense that there is a one-to-one correspondence of the readings of the MRS to the configurations of the dominance graph. However, the solvers in Sect. 3 compute *solved forms* and not configurations, and as we have noted in Sect. 2, dominance graphs may have solved forms but no

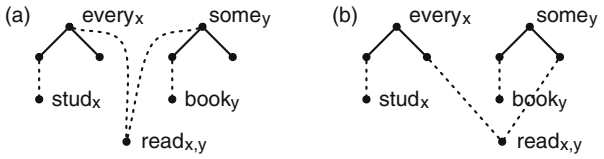


Fig. 11 (**a**) A weakly normal dominance constraint obtained from (5); (**b**) the normalizationof (**a**)

[1] Expressions $some_x(P, Q)$ and $every_x(P, Q)$ can be thought of as abbreviations for $\exists x(P \wedge Q)$ and $\forall x(P \rightarrow Q)$, respectively.

[2] The precise definition of handle constraints is slightly more restrictive, but the details are not important here.

[3] The correspondence holds only if we assume that every argument handle gets plugged by exactly one label. In general, it is possible to plug two distinct labels into the same argument handle, but this difference can be worked around [19] and we will not consider it here.

configurations. We will thus require that the graph that results from the translation must be hypernormally connected. If the graph were also normal, this would allow us to compute the pluggings of the MRS by running one of the solvers from Sect. 3, because every solved form of such a graph has a unique configuration. But due to the way variable binding is modeled in MRS, the resulting dominance graphs are usually only *weakly* normal. So we replace the dominance edges in a graph such as Fig. 11a by dominance edges that go out of the open hole of the respective fragment; the result is a normal, hypernormally connected graph as in Fig. 11b, and now we know that each of its solved forms has a configuration, each of which corresponds to exactly one plugging of the original MRS.

The crucial problem is now how to guarantee that this *normalization* step does not change the set of configurations. It can be shown that this can be done for *dominance nets* [35, 44], i.e., graphs in which (a) the height of every tree fragment is 1 or 0, (b) every tree fragment in the graph has precisely one node without outgoing dominance edges, and (c) if a node n has two or more outgoing dominance edges in G, then all its dominance children are connected by a simple hypernormal path in $G - n$. Taking all these together, we obtain the result that MRSs can be translated into normal dominance graphs in linear time, and if the graph is a dominance net, then the pluggings of the MRS bijectively correspond to the minimal solved forms of the graph.

4.2 Experiments with the English Resource Grammar

Beyond the contribution to theoretical clarity, the translation has a considerable potential for NLP applications: They make large-scale grammars that compute MRS descriptions available to users of dominance graphs, and allow users of MRS to use the efficient solvers from Sect. 3 and the disambiguation algorithms discussed below. However, for the translations to be useful, we must show that the restrictions assumed by the theorems are actually satisfied. In particular, we have to show that all MRS descriptions that are derived by grammars are actually dominance nets.

For small grammars, such a claim can be proved formally [24]. For larger grammars this is not feasible. However, we can still evaluate the claim empirically by checking what percentage of MRSs that a grammar derives for a given corpus of input sentences are dominance nets. Below, we report on experiments that we have done with the ERG, a wide-coverage grammar for English that can be seen as the reference grammar for MRS. It turns out that the majority of MRSs that the ERG computes for a large set of ordinary sentences of English are actually nets, but there is also a substantial number of MRSs which are not. Closer inspection reveals that virtually all MRSs that are intuitively correct are nets; non-nets seem to be systematically incomplete, and the Net Hypothesis can thus be turned around and made into a tool for grammar debugging.

4.2.1 Evaluating the Net Hypothesis

In order to evaluate the Net Hypothesis empirically, we considered the Rondane and Redwoods treebanks. The Redwoods treebank consists of about 10,000 sentences from dialogues in an appointment scheduling domain [36], and the Rondane corpus consists of about 1,000 sentences from the tourism domain. Each sentence in these treebanks is annotated with the preferred syntactic analysis computed by the ERG. The grammar assigns each sentence a unique MRS.

If we consider the MRS descriptions for the analyses in the October 2004 version of the treebanks, we can observe that between 72.3% (Redwoods) and 81.4% (Rondane) of the MRS descriptions that the ERG assigns to the annotated syntactic analyses are indeed nets [19]. These numbers show that a large majority of the MRS descriptions generated by the ERG are indeed nets, but there is also a substantial class of MRSs that are not nets. However, upon closer inspection, these non-nets seem to be systematically incomplete: They are missing dominance edges that should intuitively be there (e.g., because an object variable and the dominance edge from its quantifier that it induces is not there), and by adding these dominance edges would become nets after all. For instance, Fig. 12 shows the MRS for the sentence "we stay in this dramatic region for two days" (Rondane 193). The MRS is not a net. At the same time, it is evident that this MRS is incomplete, as there is no connection between the left quantifier and the "for" predicate. A second indicator for this incompleteness is that the average number of readings of non-nets grows much faster in the size of the USR if compared to the average number of nets: For instance, non-nets with 12 fragments have five times more readings than nets of the same size; for non-nets with 15 fragments, the factor is already ten.

4.2.2 Grammar Verification

So let us turn the Net Hypothesis around: Assuming that all MRSs in the treebanks *should* be nets, those 20–30 % of MRSs that are not nets must come from incorrect rules in the ERG. This is not an unreasonable assumption, given that the syntax-semantics interface is the most complex component of the grammar, which accounts

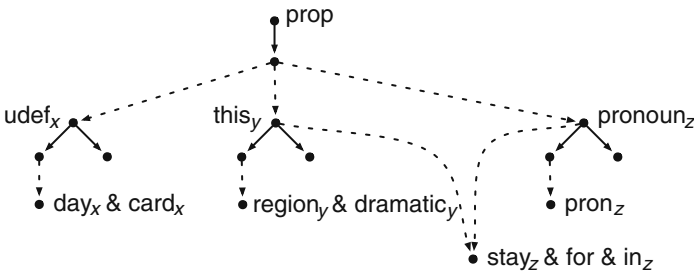


Fig. 12 MRS for "we stay in this dramatic region for two days" (Rondane 193)

for approx. 90% of the development time [10], is hard to debug, and therefore is prone to errors.

In an experiment we performed together with Dan Flickinger, one of the main developers of the ERG, we considered a rule in the ERG to be a candidate for being incorrect if all MRSs in the Rondane treebank in whose derivation it was involved were non-nets [17]. There were 11 such rules. Upon manual inspection, all 11 turned out to be incorrect. By correcting the rules, the percentage of nets among all MRSs in Rondane rose from 81 to 90%; even more encouragingly, the percentage of ill-formed MRSs (containing unbound variables, cycles, etc.) dropped from 8 to 2%, which is a clear indicator that the changes to the grammar captured true bugs, and we did not simply coerce the ERG into fitting an artificial hypothesis. Partly as a result of this new grammar debugging method, the most recent version of the ERG grammar (November 2006) now derives about 97% nets on the Rondane treebank.

In summary, the empirical evaluation of the Net Hypothesis on HPSG-annotated treebanks provides strong support for the claim that the hypothesis is correct. In fact, by assuming it is true and using it to identify trouble spots in the grammar, a grammar developer can fix bugs that had nothing to do with nets in the first place, and thus improve the quality of the syntax-semantics interface. There is a single type of MRS that we are aware of which is simultaneously linguistically valid and not a net (Copestake and Lascarides, Personal Communication). But structures of this kind are extremely rare – we found only one MRS in the two treebanks that has the same problematic structure – and could potentially be captured by a more general definition of dominance nets.

4.3 Redundancy Elimination

The central motivation of scope underspecification has always been to not simply derive an underspecified representation and then compute the readings it represents, but to perform some useful operations on the level of the underspecified representations. In the early days of underspecification, "useful operations" meant performing inferences to derive new information before disambiguating. For instance, if we know that Peter is a man, then the sentence "every man loves a woman" implies that Peter loves a woman, regardless of the particular scope reading of the sentence. Research on this problem of *direct deduction* was always hampered by the difficulty of defining what, exactly, entailment between underspecified descriptions should mean [45, 23].

Instead of direct deduction, research in the CHORUS project focused on using inferences to reduce the set of readings described by an USR. One problem in dealing with scope ambiguity is that the semantics construction process often computes an underspecified description that has multiple readings which are all semantically equivalent to each other. For instance, the sentence "a man loves a woman" has two readings which differ only in the relative scope of the two indefinites. Semantically, these two readings are equivalent, and in the context of an application, it would be sufficient to consider only one of them. In an underspecification context, we can

thus look at the *redundancy elimination* problem [46]: Given an USR U, compute another USR U' that describes a proper subset of U's readings, in such a way that for every reading φ that is described by U but not U', there is another reading that is still described by U' and semantically equivalent to it. Ideally, we would like to reduce the set of readings in such a way that no two readings are equivalent any longer, but even some reduction is useful.

Spurious ambiguities are very frequent in practice, especially when using the ERG. Consider the following two sentences from the Rondane corpus:

(6) For travelers going to Finnmark there is a bus service from Oslo to Alta through Sweden (Rondane 1262).
(7) We quickly put up the tents in the lee of a small hillside and cook for the first time in the open (Rondane 892).

For the annotated syntactic analysis of (6), the ERG derives an USR with eight scope bearing operators, which results in a total of 3,960 readings. These readings are all semantically equivalent to each other. On the other hand, the USR for (7) has 480 readings, which fall into two classes of mutually equivalent readings, characterized by the relative scope of "the lee of" and "a small hillside." Thus a redundancy elimination algorithm would be useful for working with the ERG.

It is possible to eliminate redundant readings based on charts as presented in Sect. 3.4 [28]. The key idea is as follows. Assume that a subgraph has two splits S_1 and S_2, and S_1 has a quantifier fragment Q at the root. Now let us say that we had a way of detecting efficiently that for any configuration of S_2, we can move Q to the root of that configuration by equivalence transformations. Then we could delete the split S_1 from the chart, because each of its configurations is equivalent to some configuration of S_2. In this way, we have reduced the set of readings the chart describes, without losing any equivalence classes. Note that the algorithm is correct only for USRs which are hypernormally connected (or nets), but as we have seen above, this restriction does not limit the applicability of the algorithm.

Let us illustrate this with an example. Figure 10a shows a dominance graph with five readings, all of which are semantically equivalent; Fig. 10b shows the chart of this graph as computed by the algorithm in Sect. 3.4. Now consider the splits for the complete subgraph with root fragments 1 and 2. Both of these fragments are existential quantifiers, so we can exchange their positions in a reading without changing the semantics. Moreover, the subgraph $\{2, 3, 5, 6, 7\}$ which contains the fragment 2 in the split for 1 contains only one other fragment which could possibly outscope 2; this is the fragment 3, and 3 is also an existential quantifier and can change positions with 2. Therefore we can transform every configuration described by the split for 2 into a configuration described by the split for 1 by "pushing" 2 into some lower position. This means that every configuration of the split for 2 is equivalent to some configuration of the split for 1, and so we can delete the split for 2. We can continue this process and also delete the split for 3. This reduces the chart to describing only a single reading, i.e., we have completely eliminated redundancy from the USR.

In general, this algorithm is not complete: There are USRs in which some redundancy remains after running the algorithm. However, we have shown that it performs very well on the Rondane treebank [28]. The algorithm reduces the median number of readings in the treebank from 56 to 4, the highest reduction factor for an individual USR being 666.240. On the other hand, the algorithm runs in negligible runtime because the redundancy elimination can be interleaved with the chart computation, and thus the overhead for detecting eliminable splits is balanced by the fact that we compute smaller charts. The algorithm has turned out to be a very useful tool for grammar developers, because it reduces the number of semantic representations which must be judged for correctness. In more recent work, we have shown how a simpler and more powerful redundancy elimination can be defined by viewing dominance charts as regular tree grammars and performing redundancy elimination by intersecting regular tree languages [26].

5 Annotating Scope

In the previous section, we have explored one mechanism by which the set of readings of an ambiguous expression can be reduced: by performing inferences which eliminate readings we are not interested in (e.g., because of redundancy). However, there is another mechanism for achieving this purpose, which is arguably at least as important in human language processing: apply a system of *preferences* to the ambiguous expression and pick one reading as the most salient one. For instance, there is a strong preference in German (unlike in English) to assign subjects scope over objects if they are in canonical word order [18]:

(8) Jeder Mann liebt eine Frau.
"Every man loves a woman."
(9) Eine Frau liebt jeder Mann.
"Every man loves a woman."

Example (9), in which the subject ("jeder Mann") and the object ("eine Frau") do not occur in the canonical word order, is perceived as ambiguous in the same way as the English translation. In (8), subject and object occur in canonical order, and there is a strong preference to assign the subject wide scope. There are further factors that induce scope preferences, such as intonation and the choice of determiners; for instance, "every" has a stronger tendency to take wide scope than "each."

Given a system of preferences, it is reasonably straightforward to compute a best reading, e.g., using algorithms for weighted regular tree grammars [26]. However, the question of obtaining preferences for scope ambiguities is largely open. A first proposal for a theory of scope preferences in German was worked out by Pafel [38]. His theory decides independently for each pair of quantifiers which should take scope over the other, and it seems straightforward to extend one of our search algorithms to choose the preferred scope order at each decision point. We chose

instead to develop a corpus-based model of scope preferences.[4] By annotating a German corpus with scope relations, we hoped to not only evaluate Pafel's model with real data, but also to obtain a novel statistical model of scope preferences.

In a first pilot study, a single annotator annotated 322 sentences from the NEGRA corpus [42] for scope. For each sentence, the annotator could mark up one or more constituents as scope-bearing elements, and then annotate scope relations between these. Scope-bearing elements were not restricted to noun phrases; adverbs, verbs, etc. were included as well. In this way, 121 sentences (37.5%) were annotated with at least one dominance relation between scope-bearing elements, illustrating that scope is not a rare phenomenon in natural language. It turned out that the annotation effort was extremely labor-intensive and required extensive discussions of difficult cases. The difficulty of this annotation style is illustrated by the following example, where even the decision which of the constituents should be annotated as a scope bearing element is hard to make.

(10) Er glaubt den Werbesprüchen nicht mehr, die ihn jedes Jahr zum Audio- oder TV-Wechsel animieren sollen.
"He no longer believes in the slogans designed to make him change his audio or TV system every year."

In a subsequent annotation effort, we simplified the annotation task and used syntactic patterns to extract two subcorpora from the NEGRA corpus consisting of sentences in which two candidate quantifiers have been automatically marked: The "VQQ" subcorpus contains 117 sentences in which two quantified noun phrases are syntactic arguments of the same verb, and the "KNP" subcorpus consists of 52 sentences in which one quantified noun phrase was syntactically nested within another. For each sentence, four annotators annotated only the scope relation between the two candidate quantifiers, with one of the relations "dominates" (outscopes), "is dominated by" (is outscoped), "one of the NPs is not a scope bearer", "disjoint" (neither quantifier outscopes the other), "truly ambiguous", and "don't know". This annotation scheme is a slight extension of the one reported in [22], which only takes the first three relations into account. We obtained a gold standard annotation by assigning a scope relation to a sentence if at least three annotators agreed to this relation. In this way 86 sentences (73.5%) in VQQ and 43 sentences (82.6%) in KNP received gold standard annotations.

As a general rule, the inter-annotator agreement even in this more restricted annotation task was not very high – the pairwise kappa measure ranges from 0.25 to 0.64, which is comparable to the inter-annotator agreement reported in [22] given that we use a slightly richer annotation scheme. This illustrates the difficulty of the annotation task. In addition, it turned out that a reimplementation of Pafel's system only predicted the gold standard annotation on 50% (VQQ) and 46.5% (KNP) of all sentences on which both it and the gold standard made statements. Despite the

[4] The previously unpublished research reported here was carried out by Markus Egg, Katrin Erk, Anna Hunecke, Marco Kuhlmann, and the authors.

small size of our data sets, this is an indicator that Pafel's system is only moderately successful in capturing scope preferences in naturally occurring text.

A closer look reveals that one reason why scope ambiguities are so hard to annotate is that they interact with other types of ambiguities. Consider the following examples:

(11) Im Bistum Fulda beispielsweise erhielten [alle Pfarreien]$_{Q1}$ [ein Schreiben des Generalvikariats]$_{Q2}$ ...
"In the diocese of Fulda, for example, [all parishes]$_{Q1}$ received [a letter from the office of the vicar-general]$_{Q2}$ ..."

(12) Und dann stand Israel still, als [acht Soldaten]$_{Q1}$ [aller Waffengattungen]$_{Q2}$ den Sarg zur Begräbnisstätte auf dem Herzl-Berg trugen ...
"And then Israel stood still as [eight soldiers]$_{Q1}$ from [all branches of the military]$_{Q2}$ carried the coffin to the burial ground on Mount Herzl ..."

Sentence (11) illustrates that scope ambiguities can interact with type/token ambiguities: If we assume that the sentence refers to a single letter type, which is characterized by its informational content, then all parishes receive the same letter (wide scope for "a letter"); if we assume that it refers to the individual physical objects, then we must assign "a letter" narrow scope. The difference between the two readings – does each parish receive a letter with the same content? – is subtle, and makes the annotation of this sentence tricky. Sentence (12) exhibits a different kind of problem. It neither means that each of the eight soldiers belongs to all branches of the military (wide scope for "eight soldiers"), nor that each branch was represented by its own delegation of eight soldiers (wide scope for "all branches"). Instead, we get a cumulative reading, which expresses that eight soldiers and all branches participated, and there was some surjective assignment that mapped soldiers to branches. In other words, although the sentence is syntactically indistinguishable from one with a scope ambiguity, it is not scopally ambiguous, but rather subject to issues of plural semantics.

In summary, our experiments with scope annotation were a learning experience from which we hope further research will be able to benefit. We were not able to achieve sufficient inter-annotator agreement to make the annotations useful. However, our data is already a strong indicator that Pafel's theory of scope is not sufficient to explain all observations in real-life corpora, and is a source of examples for the interaction with other phenomena such as type-token ambiguity and plural.

6 Conclusion

In this chapter, we outlined the results of the CHORUS project on the topic of semantic underspecification based on dominance constraints and dominance graphs, along two major lines of research. First, we presented a series of solvers for underspecified representations for scope ambiguities, which went hand in hand with a deeper insight into the nature and structural properties of natural language scope

underspecification. In a second, orthogonal line of research, we developed the concept of hypernormally connected dominance graphs to obtain wide-coverage semantic construction (based on translating MRSs computed by large-scale HPSG grammars) and explore methods for direct reasoning with underspecified representations.

Taken together, we ended up with a clear view of the linguistically relevant fragment of dominance-based tree descriptions and of the formal relationships between different underspecification formalisms. These theoretical insights allowed us to develop the fastest known solvers for underspecified representations, and the Net Hypothesis (every underspecified representation that is used in practice translates to a hypernormally connected dominance graph) carries far enough that it can be used for debugging grammars. Thus the research we have summarized in this chapter has contributed significantly to both the theory and the practice of processing scope ambiguities.

We have implemented the key algorithms described in this chapter and are making them available in Utool (the Swiss Army Knife of Underspecification). Utool is an open-source Java program, and it is available online at http://www.coli.uni-saarland.de/projects/chorus/utool/. Next to solving dominance graphs efficiently and performing redundancy elimination, Utool can convert between dominance graphs and a number of other underspecification formalisms, including MRS and Hole Semantics, and visualize underspecified representations. It is used by a number of grammar developers and distributed, e.g., with the GRAMMIX grammar development environment [31].

Nevertheless, there is a number of research questions that remain open. While we have shown how dominance graphs can be obtained from HPSG and dependency grammars [12], it would be interesting to investigate semantic construction from other grammar formalisms, such as TAG and LFG. Furthermore, the issue of direct deduction, of which we have solved one special case as a proof of concept, deserves further attention. Underspecification based on tree grammars and tree automata, which is compatible with dominance-based approaches through the dominance chart solver, might provide new methods for tackling both problems. Finally, with respect to the question of modeling and estimating scope preferences, we could only clarify some of the challenges related to annotating scope; solving this issue remains future research as well.

CHORUS as a long-term project. Since this chapter is written at the end of a long project, let us conclude with a few comments of what we think made CHORUS so successful. First of all, we have to mention the substantial contributions of a number of researchers who do not show up as authors of this chapter, but who either had positions in the project, or collaborated from outside. We would like to explicitly express our thanks to Ralph Debusmann, Denys Duchier, Markus Egg, Katrin Erk, Marco Kuhlmann, and Peter Ruhrberg; Etienne Ailloud, Manuel Bodirsky, Ruth Fuchss, Andrea Heyl, Anna Hunecke, Sebastian Miele, Sebastian Pado, Michaela Regneri, Kristina Striegnitz, and Feiyu Xu; and Ernst Althaus, Dan Flickinger, Michael Kohlhase, Kurt Mehlhorn, Sven Thiel, and Ralf Treinen. Most of all, our thanks go to the key players in the project from the Computer Science Department, Gert Smolka and Joachim Niehren.

Second, the project thrived on the interdisciplinarity of the SFB 378. CHORUS consistently pursued the policy of identifying fragments of representation formalisms that provide an optimal tradeoff between linguistically relevant expressive power and processing efficiency. The project started out from solid expertise in (computational) linguistics and computer science, and it provided a training ground for the continuous improvement of the participating researchers' sense of the conditions and methods of the complementary subject. CHORUS also benefited from collaborations with other projects within the SFB, and some results heavily rely on research efforts done in other projects. In particular, the development of the early dominance constraint solvers was only possible due to the Oz programming environment developed in the SFB (Project NEP).

Third, the reported results could only be achieved because the DFG funding scheme of Collaborative Research Centers and our reviewers gave us the exceptional opportunity to conduct a very long-term project, spanning a period of 12 years in total, with four subsequent funding periods. The first 3-year period was largely spent on looking for the precise formulation of the project's research question in an appropriate formal framework. This was done in a trial-and-error process, which included the exploration of higher-order and context unification to model the interaction between scope underspecification and ellipsis, the invention of lambda-structures to solve the capturing problem in reconstruction processes, and the choice of dominance constraints as a sustainable framework for the modeling of scope underspecification. Research in the subsequent funding periods focused on the investigation of more and more efficient processing techniques for dominance constraints described in the main part of this chapter, including the switch to graph algorithms as a processing paradigm. Moreover, we explored the applicability of the underspecification framework to related tasks like ellipsis reconstruction, direct inference on underspecified representation, and the modeling of metonymic reinterpretation processes. We had the chance to continuously refine and revise our formal framework – from context unification via CLLS, dominance constraints, and dominance graphs to regular tree grammars – and redesign our demo implementation into Utool, a stable system that continues to find users beyond the project itself.

The final project phase brought a breakthrough in several respects concerning practical application. The most prominent one is the specification, implementation, and empirical evaluation of an interface to a large "real world" grammar, as described in Sect. 4. The syntax-semantics interface had been on the agenda of the CHORUS project from the very beginning, and we provided a proof of concept grammar early on covering the most relevant syntactic constructions. However, we resisted suggestions from reviewers and sister projects to work out an interface to an existing large grammatical framework. In retrospect, our decision – wait with the development of a wide-coverage interface until we gained a deeper understanding of our formal framework and a clear view of its relation to the semantic representations used elsewhere – has turned out to be exactly right. One lesson for new projects we derive from this is therefore: Listen carefully to the advice of colleagues and reviewers, but at the same time deliberate carefully in what manner and on what schedule you want to follow them.

References

1. Althaus, E., Duchier, D., Koller, A., Mehlhorn, K., Niehren, J., Thiel, S. An efficient graph algorithm for dominance constraints. Journal of Algorithms, 48:194–219 (2003).
2. Apt, K.R. Principles of Constraint Programming. Cambridge: Cambridge University Press (2003).
3. Backofen, R., Rogers, J., Vijay-Shanker, K. A first-order axiomatization of the theory of finite trees. Journal of Logic, Language, and Information, 4:5–39 (1995).
4. Bender, E., Flickinger, D.P., Oepen, S. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia (2002).
5. Bodirsky, M., Duchier, D., Niehren, J., Miele, S. An efficient algorithm for weakly normal dominance constraints. In: ACM-SIAM Symposium on Discrete Algorithms. The ACM Press, New York (2004).
6. Bos, J. Predicate logic unplugged. In: Amsterdam Colloquium (pp. 133–143). Amsterdam (1996).
7. Comon, H., Dauchet, M., Gilleron, R., Löding, C., Jacquemard, F., Lugiez, D., Tison, S., Tommasi, M. Tree Automata Techniques and Applications. Available at http://www.grappa.univ-lille3.fr/tata (2007).
8. Copestake, A., Flickinger, D. An open-source grammar development environment and broad-coverage english grammar using HPSG. In: Conference on Language Resources and Evaluation. The LKB system is available at http://www.delph-in.net/lkb/ (2000).
9. Copestake, A., Flickinger, D., Pollard, C., Sag, I. Minimal recursion semantics: An introduction. Journal of Research on Language and Computation, 3(2–3):281–332 (2004).
10. Copestake, A., Lascarides, A., Flickinger, D. An algebra for semantic construction in constraint-based grammars. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (pp. 132–139). Toulouse, France (2001).
11. Dalrymple, M. (Ed.). Semantics and syntax in Lexical Functional Grammar. Cambridge, MA: MIT Press (1999).
12. Debusmann, R., Duchier, D., Koller, A., Kuhlmann, M., Smolka, G., Thater, S. A relational syntax-semantics interface based on dependency grammar. In: Proceedings of the 20th COLING. Geneva (2004).
13. Duchier, D., Niehren, J. Dominance constraints with set operators. In: Proceedings of the First International Conference on Computational Logic, no. 1861 in Lecture Notes in Computer Science (pp. 326–341). Springer-Verlag, Berlin (2000).
14. Egg, M. Flexible Semantics for Reinterpretation Phenomena. Stanford, CA: CSLI Press (2005).
15. Egg, M., Koller, A., Niehren, J. The Constraint Language for Lambda Structures. Logic, Language, and Information, 10:457–485 (2001).
16. Erk, K., Koller, A., Niehren, J. Processing underspecified semantic descriptions in the constraint language for lambda structures. Research on Language and Computation, 1:127–169 (2003).
17. Flickinger, D., Koller, A., Thater, S. A new well-formedness criterion for semantics debugging. In S. Müller (Ed.), The Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar (pp. 129–142). Stanford, CA: CSLI Publications (2005).
18. Frey, W. Syntaktische bedingungen für die semantisch interpretation. über bindung, implizite argumente und skopus. studia grammatika (1993).
19. Fuchss, R., Koller, A., Niehren, J., Thater, S. Minimal recursion semantics as dominance constraints: Translation, evaluation, and analysis. In: Proceedings of the 42nd ACL. Barcelona (2004).
20. Gardent, C., Webber, B. Describing discourse semantics. In: Proceedings of the 4th TAG+ Workshop. Philadelphia (1998).

21. Gottlob, G., Koch, C., Schulz, K.U. Conjunctive queries over trees. In: 23rd ACM Symposium on Principles of Database Systems (pp. 189–200). ACM Press, New York (2004).
22. Higgins, D., Sadock, J. A machine learning approach to modeling scope preferences. Computational Linguistics, 29(1):73–96 (2003).
23. Jaspars, J., Koller, A. A calculus for direct deduction with dominance constraints. In: Proceedings of the 12th Amsterdam Colloquium. Amsterdam (1999).
24. Koller, A., Niehren, J., Thater, S. Bridging the gap between underspecification formalisms: Hole semantics as dominance constraints. In: EACL'03. Budapest (2003).
25. Koller, A., Niehren, J., Treinen, R. Dominance constraints: Algorithms and complexity. In: Proceedings of LACL (pp. 106–125). Appeared in 2001 as volume 2014 of LNAI, Springer Verlag (1998).
26. Koller, A., Regneri, M., Thater, S. Regular tree grammars as a formalism for scope underspecification. In: Proceedings of ACL-08 (2008).
27. Koller, A., Thater, S. The evolution of dominance constraint solvers. In: Proceedings of the ACL-05 Workshop on Software. Ann Arbor (2005).
28. Koller, A., Thater, S. An improved redundancy elimination algorithm for underspecified descriptions. In: Proceedings of COLING/ACL-2006. Sydney (2006).
29. Koller, A., Thater, S. Solving unrestricted dominance graphs. In: Proceedings of the 12th Conference on Formal Grammar. Dublin (2007).
30. Marcus, M.P., Hindle, D., Fleck, M.M. D-theory: Talking about talking about trees. In: 21st Annual Meeting of the ACL (pp. 129–136). Cambridge (1983).
31. Müller, S. The Grammix CD Rom. a software collection for developing typed feature structure grammars. In T.H. King, E.M. Bender (Eds.), Grammar Engineering Across Frameworks 2007, Studies in Computational Linguistics ONLINE (pp. 259–266). Stanford, CA: CSLI Publications (2007). URL http://hpsg.fu-berlin.de/∼stefan/Pub/grammix.html
32. Müller, T., Müller, M. Finite set constraints in Oz. In F. cois Bry, B. Freitag, D. Seipel (Eds.), 13. Workshop Logische Programmierung (pp. 104–115). Germany: Technische Universität München (1997).
33. Niehren, J., Koller, A. Dominance constraints in context unification. In: Proceedings of the Third Conference on Logical Aspects of Computational Linguistics (LACL '98). Grenoble, France (1998).
34. Niehren, J., Pinkal, M., Ruhrberg, P. A uniform approach to underspecification and parallelism. In: Proceedings ACL'97 (pp. 410–417). Madrid (1997).
35. Niehren, J., Thater, S. Bridging the gap between underspecification formalisms: Minimal recursion semantics as dominance constraints. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (2003).
36. Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D., Brants, T. The LinGO Redwoods treebank: Motivation and preliminary applications. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING'02) (pp. 1253–1257) (2002)
37. Oz Development Team: The Mozart Programming System. Web pages (2004). http://www.mozart-oz.org
38. Pafel, J. Skopus und logische Struktur: Studien zum Quantorenskopus im Deutschen. Habilitationsschrift, Eberhard-Karls-Universität Tübingen (1997).
39. Poesio, M. Ambiguity, Underspecification and Discourse Interpretation. In: Proceedings of the First International Workshop on Computational Semantics (1994).
40. Reyle, U. Dealing with ambiguities by underspecification: Construction, representation and deduction. Journal of Semantics, 10(1):123–179 (1993).
41. Rogers, J., Vijay-Shanker, K. Obtaining trees from their descriptions: An application to tree-adjoining grammars. Computational Intelligence, 10:401–421 (1994).
42. Skut, W., Krenn, B., Brants, T., Uszkoreit, H. An annotation scheme for free word order languages. In: Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97. Washington, DC (1997).

43. Thatcher, J.W., Wright, J.B. Generalized finite automata theory with anapplication to a decision problem of second-order logic. Mathematical Systems Theory, 2(1):57–81 (1967).
44. Thater, S. Minimal recursion semantics as dominance constraints: Graph-theoretic foundation and application to grammar engineering. Ph.D. thesis, Department of Computational Linguistics, Saarland University (2007).
45. van Deemter, K. Towards a logic of ambiguous expressions. In: K. van Deemter, S. Peters (Eds.), Semantic Ambiguity and Underspecification (pp. 203–237). Stanford, CA: CSLI Press (1996).
46. Vestre, E. An algorithm for generating non-redundant quantifier scopings. In: Proceedings of the Fifth EACL. Berlin (1991).

Dependency Grammar: Classification and Exploration

Ralph Debusmann and Marco Kuhlmann

1 Introduction

Syntactic representations based on word-to-word dependencies have a long tradition in descriptive linguistics [29]. In recent years, they have also become increasingly used in computational tasks, such as information extraction [5], machine translation [43], and parsing [42]. Among the purported advantages of dependency over phrase structure representations are conciseness, intuitive appeal, and closeness to semantic representations such as predicate-argument structures. On the more practical side, dependency representations are attractive due to the increasing availability of large corpora of dependency analyses, such as the Prague Dependency Treebank [19].

The recent interest in dependency representations has revealed several gaps in the research on grammar formalisms based on these representations: First, while several linguistic theories of dependency grammars exist (examples are Functional Generative Description [48], Meaning-Text Theory [36], and Word Grammar [24]), there are few results on their formal properties – in particular, it is not clear how they can be related to the more well-known phrase structure-based formalisms. Second, few dependency grammars have been implemented in practical systems, and no tools for the development and exploration of new grammars are available.

In this chapter, we present results from two strands of research on dependency grammar that addresses the above issues. The aims of this research were to classify dependency grammars in terms of their generative capacity and parsing complexity, and to systematically explore their expressive power in the context of a practical system for grammar development and parsing. Our classificatory results provide fundamental insights into the relation between dependency grammars and phrase structure grammars. Our exploratory work shows how dependency-based representations can be used to model the complex interactions between different dimensions of linguistic description, such as word order, quantifier scope, and information structure.

M. Kuhlmann (✉)
Department of Linguistics and Philology, Uppsala University, 75126 Uppsala, Sweden
e-mail: marco.kuhlmann@lingfil.uu.se

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_16,

Structure of the chapter. The remainder of this chapter is structured as follows. In Sect. 2, we introduce dependency structures as the objects of description in dependency grammar, and identify three classes of such structures that are particularly relevant for practical applications. We then show how dependency structures can be related to phrase structure-based formalisms via the concept of lexicalization (Sect. 3). Section 4 introduces Extensible Dependency Grammar (XDG), a metagrammatical framework designed to facilitate the development of novel dependency grammars. In Sect. 5, we apply XDG to obtain an elegant model of complex word order phenomena, in Sect. 6 we develop a relational syntax–semantics interface, and in Sect. 7 we present an XDG model of *regular dependency grammars*. We apply the ideas behind this modeling in Sect. 8, where we introduce the grammar development environment for XDG and investigate its practical utility with an experiment on large-scale parsing. Section 9 concludes the chapter.

2 Dependency Structures

The basic assumptions behind the notion of dependency are summarized in the following sentences from the seminal work of Tesnière [51]:

> The sentence is an *organized whole*; its constituent parts are the *words*. Every word that functions as part of a sentence is no longer isolated as in the dictionary: the mind perceives *connections* between the word and its neighbours; the totality of these connections forms the scaffolding of the sentence. The structural connections establish relations of *dependency* among the words. Each such connection in principle links a *superior* term and an *inferior* term. The superior term receives the name *governor* (*régissant*); the inferior term receives the name *dependent* (*subordonné*). (chap. 1, §§ 2–4; chap. 2, §§ 1–2)

We can represent the dependency relations among the words of a sentence as a graph. More specifically, the *dependency structure* for a sentence $w = w_1 \cdots w_n$ is the directed graph on the set of positions of w that contains an edge $i \rightarrow j$ if and only if the word w_j depends on the word w_i. In this way, just like strings and parse trees, dependency structures can capture information about certain aspects of the linguistic structure of a sentence. As an example, consider Fig. 1. In this graph, the edge between the word *likes* and the word *Dan* encodes the syntactic information that *Dan* is the subject of *likes*. When visualizing dependency structures, we represent (occurrences of) words by circles and dependencies among them by arrows: the source of an arrow marks the governor of the corresponding dependency and the target marks the dependent. Furthermore, following Hays [21], we use dotted lines to indicate the left-to-right ordering of the words in the sentence.

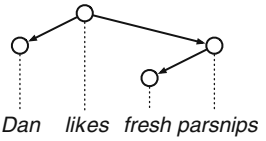


Fig. 1 A dependency structure

With the concept of a dependency structure at hand, we can express linguistic universals in terms of *structural constraints* on graphs. The most widely used such constraint is to require the dependency structure to form a tree. This requirement models the stipulations that no word should depend on itself, not even transitively, that each word should have at most one governor, and that a dependency analysis should cover all the words in the sentence. The dependency analysis shown in Fig. 1 satisfies the treeness constraint.

Another well-known constraint on dependency structures is *projectivity* [34]. In contrast to treeness, which imposes restrictions on dependency as such, projectivity concerns the relation between dependency and the left-to-right order of the words in the sentence. Specifically, it requires each dependency subtree to cover a contiguous region of the sentence. As an example, consider the dependency structure in Fig. 2a. Projectivity is interesting because the close relation between dependency and word order that it enforces can be exploited in parsing algorithms [17]. However, in recent literature, there is a growing interest in *non-projective* dependency structures, in which a subtree may be spread out over a discontinuous region of the sentence. Such representations naturally arise in the syntactic analysis of linguistic phenomena such as extraction, topicalization, a and extraposition; they are particularly frequent in the analysis of languages with flexible word order, such as Czech [22, 52]. Unfortunately, without any further restrictions, non-projective dependency parsing is intractable [40, 35].

In search of a balance between the benefit of more expressivity and the penalty of increased processing complexity, several authors have proposed structural constraints that relax the projectivity restriction, but at the same time ensure that the resulting classes of structures are computationally well-behaved [56, 41, 20]. Such constraints identify classes of what we may call *mildly non-projective dependency structures*. The *block-degree restriction* [22] relaxes projectivity such that dependency subtrees can be distributed over more than one interval. For example, in Fig. 2b, each of the marked subtrees spans two intervals. The third structural constraint that we have investigated is original to our research: *well-nestedness* [4] is the restriction that pairs of disjoint dependency subtrees must not cross, which means that there must not be nodes i_1, i_2 in the first subtree and nodes j_1, j_2 in the second such that $i_1 < j_1 < i_2 < j_2$. The dependency structure depicted in Fig. 2c is well-nested, while the structure depicted in Fig. 2b is not.

To investigate the practical relevance of the three structural constraints, we did an empirical evaluation on two versions of the Prague Dependency Treebank [33]

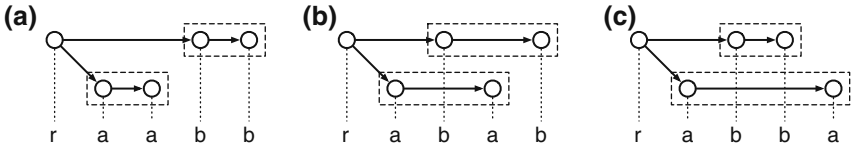


Fig. 2 Structural constraints on dependency trees

Table 1 Structural properties of dependency structures in the Prague Dependency Treebank

	PDT 1.0				PDT 2.0			
Block-degree	Unrestricted		Well-nested		Unrestricted		Well-nested	
1 (projective)	56,168	76.85%	56,168	76.85%	52,805	77.02%	52,805	77.02%
2	16,608	22.72%	16,539	22.63%	15,467	22.56%	15,393	22.45%
3	307	0.42%	300	0.41%	288	0.42%	282	0.41%
4	4	0.01%	2	< 0.01%	2	< 0.01%	–	–
5	1	< 0.01%	1	< 0.01%	1	< 0.01%	1	< 0.01%
TOTAL	73,088	100.00%	73,010	99.89%	68,562	100.00%	68,481	99.88%

(Table 1). This evaluation shows that while projectivity is too strong a constraint on dependency structures (it excludes almost 23% of the analyses in both versions of the treebank), already a small step beyond projectivity covers virtually all of the data. In particular, even the rather restricted class of well-nested dependency structures with block-degree at most 2 has a coverage of almost 99.5%.

3 Dependency Structures and Lexicalized Grammars

One of the fundamental questions that we can ask about a grammar formalism is whether it adequately models natural language. We can answer this question by studying the *generative capacity* of the formalism: when we interpret grammars as generators of sets of linguistic structures (such as strings, parse trees, or predicate-argument structures), then we can call a grammar adequate, if it generates exactly those structures that we consider relevant for the description of natural language. Grammars may be adequate with respect to one type of expression, but inadequate with respect to another. Here we are interested in the generative capacity of grammars when we interpret them as generators for sets of dependency structures:

Which grammars generate which sets of dependency structures?

An answer to this question is interesting for at least two reasons. First, dependency structures make an attractive measure of the generative capacity of a grammar: they are more informative than strings, but less formalism-specific and arguably closer to a semantic representation than parse trees. Second, an answer to the question allows us to tap the rich resource of formal results about generative grammar formalisms and to transfer them to the work on dependency grammar. Specifically, it enables us to import the expertise in developing parsing algorithms for lexicalized grammar formalisms. This can help us identify the polynomial fragments of non-projective dependency parsing.

3.1 Lexicalized Grammars Induce Dependency Structures

In order to relate grammar formalisms and dependency representations, we first need to specify in what sense we can consider a grammar as a generator of

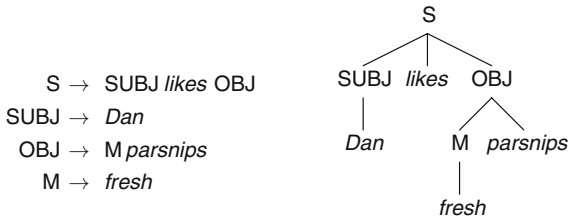


Fig. 3 A context-free grammar and a parse tree generated by this grammar

dependency structures. To focus our discussion, let us consider the well-known case of context-free grammars (CFGs). As our running example, Fig. 3 shows a small CFG together with a parse tree for a simple English sentence.

An interesting property of our sample grammar is that it is *lexicalized*: every rule of the grammar contains exactly one terminal symbol. Lexicalized grammars play a significant role in contemporary linguistic theories and practical applications. Crucially for us, every such grammar can be understood as a generator for sets of dependency structures, in the following sense. Consider a derivation of a terminal string by means of a context-free grammar. A *derivation tree* for this derivation is a tree in which the nodes are labelled with (occurrences of) the productions used in the derivation, and the edges indicate how these productions were combined. The left half of Fig. 4 shows the derivation tree for the parse tree from our example. If the underlying grammar is lexicalized, then there is a one-to-one correspondence between the nodes in the derivation tree and the positions in the derived string: each occurrence of a production participating in the derivation contributes exactly one terminal symbol to this string. If we order the nodes of the derivation tree according to the string positions of their corresponding terminal symbols, we get a dependency tree. For our example, this procedure results in the tree depicted in Fig. 1. We say that this dependency structure is *induced* by the derivation d.

Not all practically relevant dependency structures can be induced by derivations in lexicalized context-free grammars. A famous counterexample is provided by the verb-argument dependencies in German and Dutch subordinate clauses: context-free grammar can only characterize the "nested" dependencies of German, but not the "cross-serial" assignments of Dutch. This observation goes along with arguments [25, 49] that certain constructions in Swiss German require grammar formalisms that adequately model these constructions to generate the so-called copy

Fig. 4 Lexicalized derivations induce dependency structures

language, which is beyond even the string-generative capacity of CFGs. If we accept this analysis, then we must conclude that context-free grammars are not adequate for the description of natural language, and that we should look out for more powerful formalisms. This conclusion is widely accepted today. Unfortunately, the first class in Chomsky's hierarchy of formal languages that *does* contain the copy language, the class of *context-sensitive languages*, also contains many languages that are considered to be beyond human capacity. Also, while CFGs can be parsed in polynomial time, parsing of context-sensitive grammars is PSPACE-complete. In search of a class of grammars that extends context-free grammar by the minimal amount of generative power that is needed to account for natural language, several so-called *mildly context-sensitive grammar formalisms* have been developed; perhaps the best-known among these is Tree Adjoining Grammar (TAG) [27]. The class of string languages generated by TAGs contains the copy language, but unlike context-sensitive grammars, TAGs can be parsed in polynomial time. More important to us than their increased string-generative capacity however is their stronger power with respect to dependency representations: derivations in (lexicalized) TAGs can induce the "cross-serial" dependencies of Dutch [26]. The principal goal of our classificatory work is to make the relations between grammars and the dependency structures that they can induce precise.

In spite of the apparent connection between the generative capacity of a grammar formalism and the structural properties of the dependency structures that this formalism can induce, there have been only few results that link the two research areas. A fundamental reason for the lack of such bridging results is that, while structural constraints on dependency structures are *internal* properties in the sense that they concern the nodes of the graph and their connections, grammars take an *external* perspective on the objects that they manipulate – the internal structure of an object is determined by the internal structure of its constituent parts and the operations that are used to combine them. An example for the difference between the two views is given by the perspectives on trees that we find in graph theory and universal algebra. In graph theory, a tree is a special graph with an internal structure that meets certain constraints; in algebra, trees are abstract objects that can be composed and decomposed using a certain set of operations. One of the central technical questions that we need to answer in order to connect grammars and structures is how classes of dependency structures can be given an algebraic structure.

3.2 The Algebraic View on Dependency Structures

In order to link structural constraints to generative grammar formalisms, we need to view dependency structures as the outcomes of compositional processes. Under this view, structural constraints do not only apply to fully specified dependency trees, but already to the composition operations by which these trees are constructed. We formalized the compositional view in two steps. In the first step, we showed that dependency structures can be encoded into terms over a certain signature of *order*

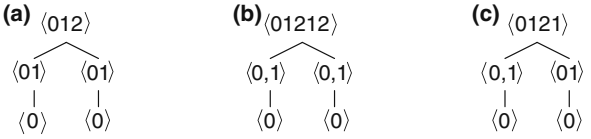


Fig. 5 (**a**) Term t_1 (**b**) Term t_2 (**c**) Term t_3

annotations in such a way that the three different classes of dependency structures that we have discussed above stand in one-to-one correspondence with terms over specific subsets of this signature [31]. In the second step, we defined the concept of a *dependency algebra*. In these algebras, order annotations are interpreted as composition operations on dependency structures [30, 32]. We have proved that each dependency algebra is isomorphic to the corresponding term algebra, which means that the composition of dependency structures can be freely simulated by the usual composition operations on terms, such as substitution.

To give an intuition for the algebraic framework, Fig. 5 shows the terms that correspond to the dependency structures in Fig. 2. Each order annotation in these terms encodes node-specific information about the linear order on the nodes. As an example, the constructor $\langle 0, 1\rangle$ in Fig. 5 represents the information that the marked subtrees in Fig. 2b each consist of two intervals (the two components of the tuple $\langle 0, 1\rangle$), with the root node (represented by the symbol 0) situated in the left interval, and the subtree rooted at the first child (represented by the symbol 1) in the right interval. Under this encoding, the block-degree measure corresponds to the maximal number of components per tuple, and the well-nestedness condition corresponds to the absence of certain "forbidden substrings" in the individual order annotations, such as the substring 1212 in the term in Fig. 5.

Our algebraic framework enables us to classify the dependency structures that are induced by various lexicalized grammar formalisms. In particular, we can extend Gaifman's result [18] that projective dependency structures correspond to lexicalized context-free grammars into the realm of the mildly context-sensitive: the classes of block-restricted dependency structures correspond to Linear Context-Free Rewriting Systems [53, 54], the classes of well-nested block-restricted structures correspond to Coupled Context-Free Grammar [23]. As a special case, the class of well-nested dependency structures with a block-degree of at most 2 is characteristic for derivations in Lexicalized Tree Adjoining Grammar [27, 4]. This result is particularly interesting in the context of our treebank evaluation.

3.3 Regular Dependency Grammars

We can now lift our results from individual dependency structures to sets of such structures. The key to this transfer is the concept of *regular sets of dependency structures* [31], which we define as the recognizable subsets of dependency algebras in the sense of Mezei and Wright [37]. Based on the isomorphism between dependency

algebras and term algebras, we obtain a natural grammar formalism for dependency structures from the concept of a *regular term grammar*.

Definition 1 A *regular dependency grammar* is a construct $G = (N, \Sigma, S, P)$, where N is a ranked alphabet of non-terminal symbols, Σ is a finite set of order annotations, $S \in N_1$ is a distinguished start symbol, and P is a finite set of productions of the form $A \to t$, where $A \in N_k$ is a non-terminal symbol, and $t \in T_{\Sigma,k}$ is a well-formed term over Σ of sort k, for some $k \in \mathbb{N}$.

To illustrate the definition, we give two examples of regular dependency grammars. The sets of dependency structures generated by these grammars mimic the verb-argument relations found in German and Dutch subordinate clauses, respectively: grammar G_1 generates structures with nested dependencies and grammar G_2 generates structures with crossing dependencies. We only give the two sets of productions.

$$S \to \langle 120\rangle(N, V) \quad V \to \langle 120\rangle(N, V) \quad V \to \langle 10\rangle(N) \quad N \to \langle 0\rangle \quad (G_1)$$
$$S \to \langle 1202\rangle(N, V) \quad V \to \langle 12, 02\rangle(N, V) \quad V \to \langle 1, 0\rangle(N) \quad N \to \langle 0\rangle \quad (G_2)$$

Figure 6 shows terms generated by these grammars and the corresponding dependency structures.

The sets of dependency structures generated by regular dependency grammars have all the characteristic properties of mildly context-sensitive languages. Furthermore, it turns out that the structural constraints that we have discussed above have direct implications for their string-generative capacity and parsing complexity. First, the block-degree measure gives rise to an infinite hierarchy of ever more powerful string languages; adding the well-nestedness constraint leads to a proper decrease of string-generative power on nearly all levels of this hierarchy [32]. Certain string languages enforce structural properties in the dependency languages that project them. For every natural number k, the language

$$COUNT(k) := \{ a_1^n b_1^n \cdots a_k^n b_k^n \mid n \in \mathbb{N} \}$$

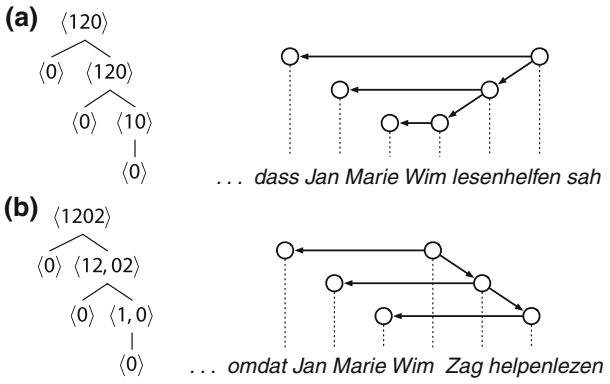


Fig. 6 (**a**) Grammar G_1 (nested dependencies). (**b**) Grammar G_2 (cross-serial dependencies)

requires every regular set of dependency structures that projects it to contain structures with a block-degree of at most k. Similarly, the language

$$RESP(k) \quad := \quad \{\, a_1^m b_1^m c_1^n d_1^n \cdots a_k^m b_k^m c_k^n d_k^n \mid m, n \in \mathbb{N} \,\}$$

requires every regular set of dependency structures with block-degree at most k that projects it to contain structures that are not well-nested. Second, while the parsing problem of regular dependency languages is polynomial in the length of the input string, the problem in which we take the grammar to be part of the input is still NP-complete. Interestingly, for well-nested dependency languages, parsing is polynomial even with the size of the grammar taken into account [30].

4 Extensible Dependency Grammar

For the *exploration* of dependency grammars, we have developed a new metagrammatical framework called Extensible Dependency Grammar (XDG) [11, 8]. The main innovation of XDG is *multi-dimensionality*: an XDG analysis consists of a tuple of dependency graphs all sharing the same set of nodes, called *dependency multigraph*. The components of the multigraph are called *dimensions*. The multidimensional metaphor was crucial for our formulations of a new, elegant model of complex word order phenomena in German, and a new, relational model of the syntax–semantics interface.

4.1 Dependency Multigraphs

To give an intuition, let us start with an example multigraph depicted in Fig. 7. The multigraph has three dimensions called DEP (for *dependency tree*), QS (for *quantifier scope analysis*), and DEP/QS (DEP/QS *syntax–semantics interface*). It is not necessary to fully understand what we intend to model with these dimensions; they just serve as an illustrative example and are elucidated in more detail in Sect. 6 below.

In an XDG multigraph, each dimension is a dependency graph made up of a set of nodes associated with indices, words, and node attributes. The indices and words are shared across all dimensions. For instance, the second node on the DEP dimension is associated with the index 2, the word *loves*, and the node attributes in, out and $order$. On the DEP/QS dimension, the node has the same index and word and the node attribute dom. Node attributes always denote sets of tuples over finite domains of atoms; their typical use is to model finite relations like functions and orders. The nodes are connected by labeled edges. On the QS dimension, for example, there is an edge from node 3 to node 1 labeled sc, and another one from node 1 to node 2, also labeled sc.

In the example, the DEP dimension states that *everybody* is the subject of *loves*, and *somebody* the object. The in and out attributes represent the licensed incoming

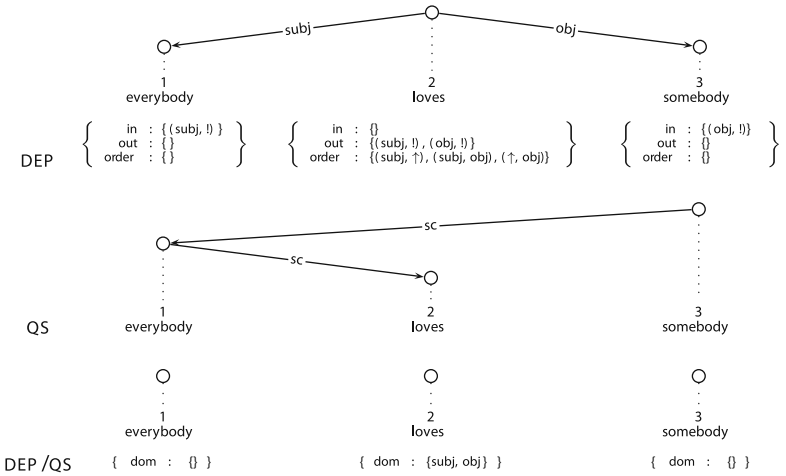


Fig. 7 Dependency multigraph for *Everybody loves somebody*

and outgoing edges. For example, node 2 must not have any incoming edges, and it must have one outgoing edge labeled subj and one labeled obj. The $order$ attribute represents a total order among the head (↑) and its dependents: the subj dependents must precede the head, and head must precede the obj dependents.

The QS dimension is an analysis of the scopal relationships of the quantifiers in the sentence. It models the reading where *somebody* takes scope over *everybody*, which in turn takes scope over *loves*. The DEP/QS analysis represents the syntax-semantics interface between DEP and QS. The attribute dom is a set of those dependents on the DEP dimension that must dominate the head on the QS dimension. For example, the subj and obj dependents of node 2 on DEP must dominate 2 on QS.

4.2 Grammars

XDG is a *model-theoretic* framework: grammars first delineate the set of all *candidate structures*, and second, all structures which are not well-formed according to a set of *constraints* are eliminated. The remaining structures are the *models* of the grammar. This contrasts with approaches such as the regular dependency grammars of Sect. 3.3, where the models are *generated* using a set of productions.

An XDG grammar $G = (MT, lex, P)$ has three components: a *multigraph type* MT, a *lexicon* lex, and a set of *principles* P. The multigraph type specifies the dimensions, words, edge labels, and node attributes and thus delineates the set of candidate structures of the grammar. The lexicon is a function from the words of the grammar to sets of lexical entries, which determine the node attributes of the nodes with that word. The principles are a set of formulas in first-order logic constituting the constraints of the grammar. Principles can talk about precedence, edges,

dominances (transitive closure[1] of the edge relation), the words associated to the nodes, and the node attributes. Here is an example principle forbidding cycles on dimension DEP. It states that no node may dominate itself:

$$\forall v : \neg(v \rightarrow^{+}_{\text{DEP}} v) \tag{1}$$

The second example principle stipulates a constraint for all edges from v to v' labeled l on dimension DEP: if l is in the set denoted by the lexical attribute dom of v on DEP/QS, then v' must dominate v on QS:

$$\forall v : \forall v' : \forall l : \; v \xrightarrow{l}_{\text{DEP}} v' \; \wedge \; l \in dom_{\text{DEP/QS}}(v) \;\; \Rightarrow \;\; v' \rightarrow^{+}_{\text{QS}} v \tag{2}$$

Observe that the principle is indeed satisfied in Fig. 7: the attribute dom for node 2 on DEP/QS includes subj and obj, and both the subj and the obj dependents of node 2 on DEP dominate node 2 on QS.

A multigraph is a *model* of a grammar $G = (MT, lex, P)$ iff it is one of the candidate structures delineated by MT, it selects precisely one lexical entry from lex for each node, and it satisfies all principles in P.

The *string language* $L(G)$ of a grammar G is the set of *yields* of its models. The *recognition problem* is the question given a grammar G and a string s, is s in $L(G)$. We have investigated the complexity of three kinds of recognition problems [9]: The *universal* recognition problem where both G and s are variable is PSPACE-complete, the *fixed* recognition problem where G is fixed and s is variable is NP-complete, and the *instance* recognition problem where the principles are fixed and the lexicon and s are variable is also NP-complete. XDG parsing is NP-complete as well.

XDG is at least as expressive as CFG [8]. We have proven that the string languages of XDG grammars are closed under union and intersection [10]. In Sect. 7, we give a constructive proof that XDG is at least as expressive as the class of regular dependency grammars introduced in Sect. 3.3, which entails through an encoding of LCFRS in regular dependency grammars, that XDG is at least as expressive as LCFRS. As XDG is able to model scrambling (see Sect. 5.2), which LCFRS is not [3], it is indeed *more* expressive than LCFRS.

5 Modeling Complex Word Order Phenomena

The first application for the multi-dimensionality of XDG in CHORUS is the design of a new, elegant model of complex word order phenomena such as *scrambling*.

[1] Transitive closures cannot be expressed in first-order logic. As in practice, the only transitive closure that we need is the transitive closure of the edge relation, we have decided to encode it in the multigraph model and thus stay in first-order logic [10].

5.1 Scrambling

In German, the word order in subordinate sentences is such that all verbs are positioned at the right end in the so-called *verb cluster* and are preceded by all the non-verbal dependents in the so-called *Mittelfeld*. Whereas the mutual order of the verbs is fixed, that of the non-verbal dependents in the Mittelfeld is totally free.[2] This leads to the phenomenon of scrambling. We show an example in Fig. 8, where the subscripts indicate the dependencies between the verbs and their arguments.

Mittelfeld	verbcluster
(dass) $Nilpferde_3$ $Maria_1$ $Hans_2$	$füttern_3$ $helfen_2$ $soll_1$
(that) $hippos_3$ $Maria_1$ $Hans_2$	$feed_3$ $help_2$ $should_1$
(that) Maria should help Hans feed hippos	

Fig. 8 Example for scrambling

In the dependency analysis in Fig. 9 (top), we can see that scrambling gives rise to *non-projectivity*. In fact, scrambling even gives rise to an unbounded block-degree (see Sect. 2), which means that it can neither be modeled by LCFRS nor by regular dependency grammars.

5.2 A Topological Model of Scrambling

As we have proven [8], scrambling *can* be modeled in XDG. But how? There is no straightforward way of articulating appropriate word order constraints on the DEP dimension directly. At this point, we can make use of the multi-dimensionality

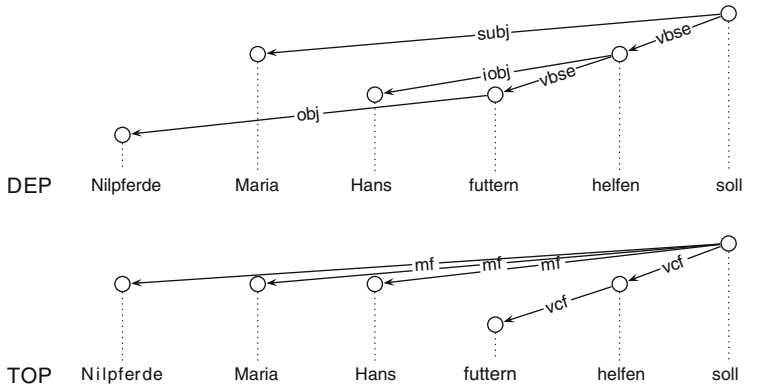


Fig. 9 Dependency analysis (*top*) and topological analysis (*bottom*) of the scrambling example

[2] These are of course simplifications: the order of the verbs can be subject to alternations such as *Oberfeldumstellung*, and although all linearizations of the non-verbal dependents are grammatical, some of them are clearly marked.

of XDG. The idea is to keep the dependency analysis on the DEP dimension as it is, and move all ordering constraints to an additional dimension called TOP. The models on TOP are *projective* trees which represent the topological structure of the sentence as in Topological Dependency Grammar (TDG) [15]. A TOP analysis of the example sentence is depicted in Fig. 9 (bottom). Here, the non-verbal dependents *Nilpferde*, *Maria*, and *Hans* are dependents of the finite verb *soll* labeled mf for "Mittelfeld". The verbal dependent of *soll*, *helfen*, and that of *helfen*, *füttern*, are labeled vcf for "verb cluster field". With this additional dimension, articulating the appropriate word order constraints is straightforward: all mf dependents of the finite verb must precede its vcf dependents, and the mutual order of the mf dependents is unconstrained.

The relation between the DEP and TOP dimensions is such that the trees on TOP are a *flattening* of the corresponding trees on DEP. We can express this in XDG by requiring that the dominance relation on TOP is a subset of the dominance relation on DEP:

$$\forall v : \forall v' : \; v \rightarrow^{+}_{\text{TOP}} v' \;\Rightarrow\; v \rightarrow^{+}_{\text{DEP}} v'$$

This principle is called the *climbing principle* [15], and gets its name from the observation that the non-verbal dependents seem to "climb up" from their position on DEP to a higher position on TOP. For example, in Fig. 9, the noun *Nilpferde* is a dependent of *füttern* on DEP, and climbs up to become a dependent of the finite verb *soll* on TOP.

Just using the climbing principle is too permissive. For example, in German, extraction of determiners and adjectives out of noun phrases must be ruled out, whereas relative clauses *can* be extracted. To this end, we apply a principle called *barriers principle* [15], which allows each word to "block" certain dependents from climbing up. This allows us to express that nouns block their determiner and adjective dependents from climbing up, but not their relative clause dependents.

6 A Relational Syntax–Semantics Interface

Our second application of XDG is the realization of a new, relational syntax-semantics interface [11]. The interface is relational in the sense that it constrains the *relation* between the syntax and the semantics, as opposed to the traditional functional approach where the semantics is *derived* from syntax. In combination with the constraint-based implementation of XDG, the main advantage of this approach is *bi-directionality*: the same grammar can be "reversed" and be used for generation, and constraints and preferences can "flow back" from the semantics to disambiguate the syntax. In this section, we introduce the subset of the full relational syntax–semantics interface for XDG [11], concerned only with the relation between grammatical functions and quantifier scope. We model quantifier scope using dominance

constraints, the integral part of the Constraint Language for Lambda Structures (CLLS) [16].

6.1 Dominance Constraints

Dominance constraints can be applied to model the *underspecification* of scopal relationships. For example, the sentence *Everybody loves somebody* has two scopal readings: one where everybody loves the same somebody, and one where everybody loves somebody else. The first reading is called the *strong reading*: here, *somebody* takes scope over *everybody*, which in turn takes scope over *loves*. In the second reading, the *weak reading*, *everybody* takes scope over *somebody* over *loves*. Using dominance constraints, it is possible to model both readings in one underspecified representation called *dominance constraint*:

$$X_1 : everybody(X'_1) \wedge X_2 : loves \wedge X_3 : somebody(X'_3) \wedge X'_1 \lhd^* X_2 \wedge X'_3 \lhd^* X_2 \qquad (3)$$

The dominance constraint comprises three *labeling literals* and two *dominance literals*. A labeling literal such as $X_1 : everybody(X'_1)$ assigns labels to node variables, and constrains the daughters: X_1 must have the label *everybody*, and it must have one daughter, viz. X'_1. The dominance literals $X'_1 \lhd^* X_2$ and $X'_3 \lhd^* X_2$ stipulate that the node variables X'_1 and X'_3 must dominate (or be equal to) the node variable X_2, expressing that the node variables corresponding to *everybody* and *somebody* must dominate *loves*, but that their mutual dominance relationship is unknown.

The models of dominance constraints are trees called *configurations*. The example dominance constraint (3) represents the two configurations displayed in Fig. 10 (a) (strong reading) and (b) (weak reading).

In XDG, we represent the configurations on a dimension called QS for *quantifier scope analysis*. For example, the configuration in Fig. 10 (a) corresponds to the XDG dependency tree in Fig. 10b, c, d. The QS dimension must satisfy only one principle: it must have tree-shape.

We model the dominance constraint itself by translating the labeling literals into constraints on the node-word mapping, and the dominance literals into dominance predicates. Hereby, we conflate the node variables participating in labeling literals

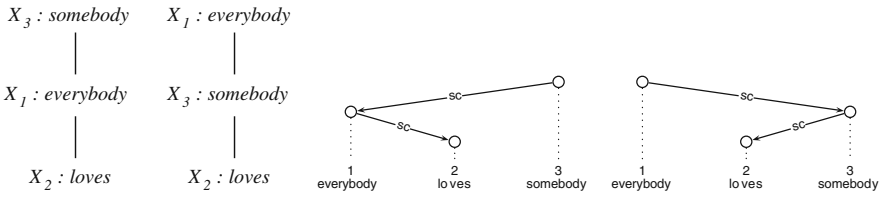


Fig. 10 (**a**) Configuration representing the strong reading, (**b**) the weak reading of *Everybody loves somebody*, (**c**) corresponding XDG dependency tree for the strong reading, (**d**) weak reading

such as $X_1 : everybody(X'_1)$ into individual nodes.[3] We translate the dominance constraint (3) into the following XDG principle:

$$w(1) = everybody \ \wedge \ w(2) = loves \ \wedge \ w(3) = somebody \ \wedge \ 1 \rightarrow^*_{\text{QS}} 2 \ \wedge \ 3 \rightarrow^*_{\text{QS}} 2 \qquad (4)$$

The set of QS tree structures which satisfy this principle corresponds precisely to the set of configurations of the dominance constraint in (3), i.e., the two dependency trees in Fig. 10c, d.

6.2 The Interface

We now apply this formalization of dominance constraints in XDG for building a relational syntax–semantics interface. The interface relates DEP, a syntactic dimension representing grammatical functions, and QS, a semantic dimension representing quantifier scope, and consists of two ingredients: the additional interface dimension DEP/QS and an additional interface principle. The models on DEP/QS have no edges, as the sole purpose of this dimension is to carry the lexical attribute dom specifying how the syntactic dependencies on DEP relate to the quantifier scope dependencies on QS. The value of dom is a set of DEP edge labels, and for each node, all syntactic dependents with a label in dom must dominate the node on QS. We call the corresponding principle, already formalized in (2), *dominance principle*.

This kind of syntax–semantics interface is "two-way", or *bi-directional*: information does not only flow from syntax to semantics, but also vice versa. Like a functional interface, given a syntactic representation, the relational interface is able to derive the corresponding semantic representation. For example, the two syntactic dependencies from node 2 (*loves*) to node 1 (labeled subj) and to node 3 (labeled obj) in Fig. 7, together with the dominance principle, yield the information that both the subject and the object of *loves* must dominate it on the QS dimension.

The relational syntax–semantics interface goes beyond the functional one in its ability to let information from the semantics "flow back" to the syntax. For example, assume that we start with a partial QS structure including the information that *everybody* and *somebody* both dominate *loves*. Together with the dominance principle, this excludes any edges from *everybody* to *loves* and from *somebody* to *loves* on DEP.[4] Thus, information from the semantics has disambiguated the syntax. This bi-directionality can also be exploited for "reversing" grammars to be used for generation as well as for parsing [28, 7].

---

[3] In our simple example, the labeling literals have at most one daughter. In a more realistic setting [8], we distinguish the daughters of labeling literals with more than one daughter using distinct edge labels.

[4] For the example, we assume that the value of dom for *everybody* and *somebody* includes all edge labels.

7 Modeling Regular Dependency Grammars

In this section, we apply XDG to obtain a multi-dimensional model of regular dependency grammars (REGDG). This not only gives us a lower bound of the expressivity of XDG, but also yields techniques for parsing TAG grammars. We demonstrate the application of the latter to large-scale parsing in Sect. 8.2.

Our modeling of REGDG proceeds in two steps. In the first, we examine the *structures* that they talk about: totally ordered dependency trees. To ease the transition to XDG, we replace the node labels in the REGDG dependency trees by edge labels in the XDG dependency trees. Figure 11 (top) shows such a dependency tree of the string *aaabbb*. We call the XDG dependency tree dimension DEP. On the DEP dimension, the models must have tree-shape but need not be projective.

In the second step, we examine the *rules* of REGDG. They can be best explained by example. Consider the rule

$$\mathrm{A} \rightarrow \langle a, \langle 01, 21 \rangle \rangle (\mathrm{A}, \mathrm{B}) \tag{5}$$

which is expanded by the second *a* in Fig. 11. First, the rule stipulates that a head with incoming edge label A associated with the word *a* must have two dependents: one labeled A and one labeled B. Second, the rule stipulates the order of the yields of the dependents and the head, where the yields are divided into contiguous sets of nodes called *blocks*. In the order tuples (e.g., $\langle 01, 21 \rangle$), 0 represents the head, 1 the blocks in the yield of the first dependent (here: A), and 2 the blocks in the yield

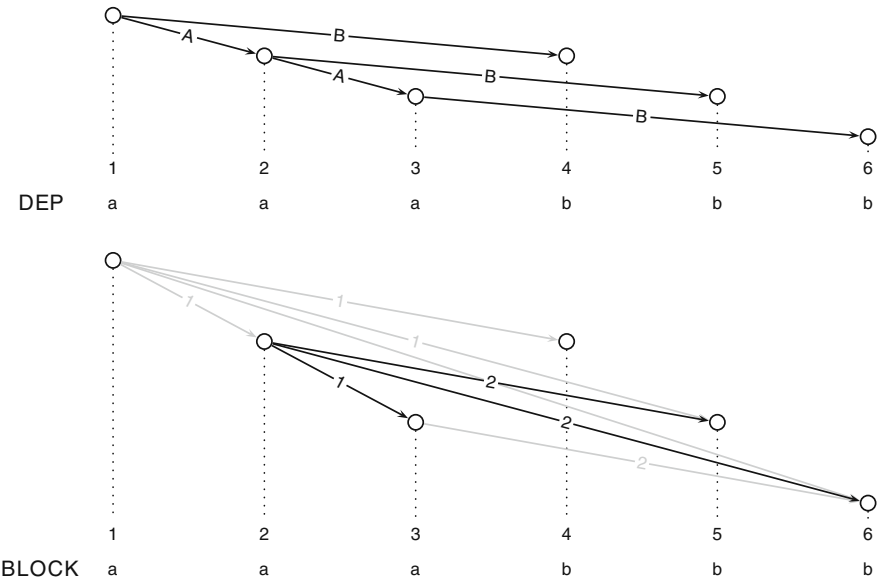


Fig. 11 XDG dependency tree (*top*) and XDG block graph (*bottom*) for the string *aaabbb*

of the second dependent (here: B). The tuple $\langle 01, 21 \rangle$ from the example rule then states: the yield of the A dependent must consist of two blocks (two occurrences of 1 in the tuple) and that of the B dependent of one block (one occurrence of 2), the head must precede the first block of the A dependent, which must precede the first (and only) block of the B dependent, which must precede the second block of the A dependent, and the yield of the head must be divided into two blocks, where the gap is between the first block of the A dependent and the first (and only) block of the B dependent.

As REGDG do not only make statements on dependency structures but also on the yields of the nodes, we exploit the multi-dimensionality of XDG and introduce a second dimension called BLOCK. The structures on the BLOCK dimension are graphs representing the function from nodes to their yields on DEP. That is, each edge from v to v' on BLOCK corresponds to a sequence of zero or more edges from v to v' on DEP:

$$\forall v : \forall v' : \; v \rightarrow_{\text{QS}} v' \;\Leftrightarrow\; v \rightarrow^{*}_{\text{DEP}} v'$$

An edge from v to v' labeled i on BLOCK states that v' is in the ith block of the yield of v on DEP.

We model that the blocks are *contiguous* sets of nodes by a principle stipulating that for all pairs of edges, one from v to v', and one from v to v'', both labeled with the same label l, the set of nodes between v' and v'' must also be in the yield of v:

$$\begin{aligned} \forall v : \forall v' : \forall v'' : \forall l : (v \xrightarrow{l}_{\text{BLOCK}} v' \wedge v \xrightarrow{l}_{\text{BLOCK}} v'') \\ \Rightarrow (\forall v''' : v' < v''' \wedge v''' < v'' \Rightarrow v \rightarrow^{*}_{\text{BLOCK}} v''') \end{aligned}$$

Figure 11 (bottom)[5] shows an example BLOCK graph complementing the DEP tree in Fig. 11 (top). On DEP, the yield of the second *a* (node 2) consists of itself and the third *a* (node 3) in the first block, and the second *b* and the third *b* (nodes 5 and 6) in the second block. Hence, on the BLOCK dimension, the node has four dependents: itself and the third *a* are dependents labeled 1, and the second *b* and the third *b* are dependents labeled 2.

We model the rules of the REGDG in XDG in four steps. First, we lexically constrain the incoming and outgoing edges of the nodes on DEP. For example, to model the example rule (5), we stipulate that the node associated with the word *a* must have precisely one incoming edge labeled A, and one A and one B dependent, as shown in Fig. 12a.

Second, we lexically constrain the incoming and outgoing edges on BLOCK. As each node on BLOCK can end up in any block of any other node, each node may have arbitrary many incoming edges either labeled 1 or 2. The constraint on the outgoing edges reflects the number of blocks into which the yield of the node must be divided. For the example rule (5), the yield must be divided into two blocks, and hence the

[5] For clarity, the graphical representation does not include the edges from each node to itself, and all edges except those emanating from node 2 are ghosted.

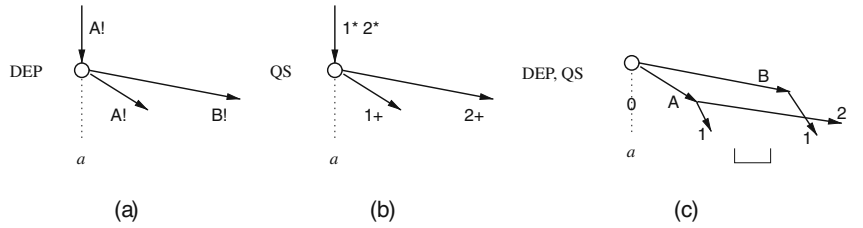


Fig. 12 (**a**) Constraints on DEP, (**b**) BLOCK, and (**c**) on both DEP and BLOCK

node must have one or more dependents labeled 1, and one or more labeled 2, as depicted in Fig. 12b.

Third, we lexically constrain the order of the blocks of the DEP dependents. We do this by a constraint relating the DEP and BLOCK dimensions. For the example rule (5), we must order the head to the left of the first block of the A dependent. In terms of our XDG model, as illustrated in Fig. 12c, we must order the head to the left of all 1 dependents on BLOCK of the A dependent on DEP. Similarly, we must order the 1 dependents of the A dependent to the left of the 1 dependents of the B dependent, and these in turn to the left of the 2 dependents of the A dependent.

Fourth, we lexically model the location of the gaps between the blocks. In the example rule (5), there is one gap between the first block of the A dependent and the first (and only) block of the B dependent, as indicated in Fig. 12c.

8 Grammar Development Environment

We have complemented the theoretical exploration of dependency grammars using XDG with the development of a comprehensive grammar development environment, the XDG Development Kit (XDK) [13, 8, 47]. The XDK includes a parser, a powerful grammar description language, an efficient compiler for it, various tools for testing and debugging, and a graphical user interface, all geared towards rapid prototyping and the verification of new ideas. It is written in MOZART/OZ [50, 38]. A snapshot of the XDK is depicted in Fig. 13.

8.1 Parser

The included parser is based on constraint programming [45], a modern technique for solving NP-complete problems. For efficient parsing, the applied constraints implementing the XDG principles must be fine-tuned. Fine-tuned implementations of the principles of the account of word order outlined in Sect. 5, and the relational syntax–semantics interface outlined in Sect. 6 already exist, and have yielded efficiently parsable, smaller-scale grammars for German [6, 2] and English [8]. Koller and Striegnitz show that an implementation of TDG can be applied for efficient TAG generation [28].

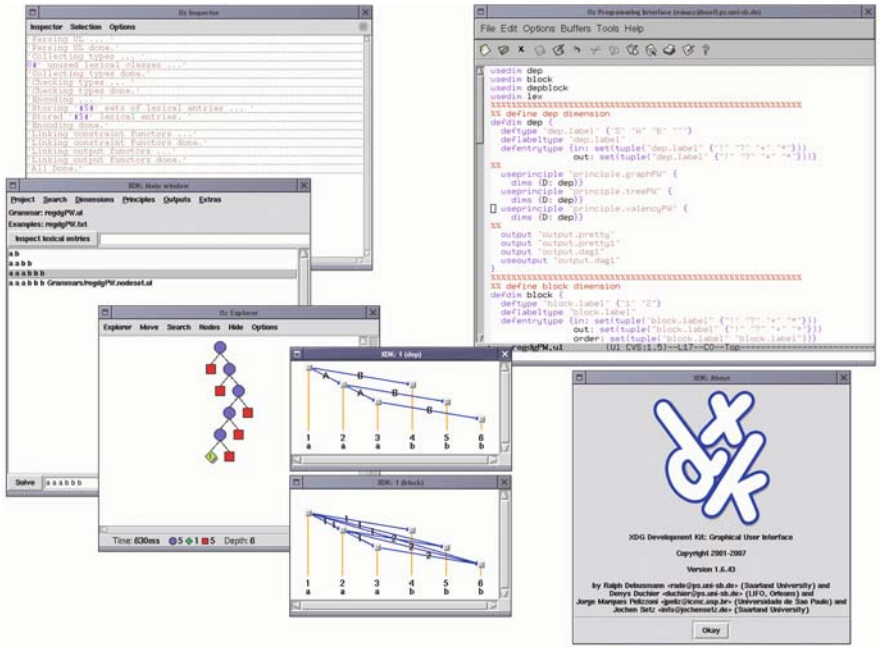


Fig. 13 The XDG Development Kit (XDK)

8.2 Large-Scale Parsing

In this section, we answer the question whether the constraint parser of the XDK actually scales up for large-scale parsing. We find a positive answer to this question by showing that the parser can be fine-tuned for parsing the large-scale TAG grammar XTAG [55], such that most of the time, it finds the first parses of a sentence *before* a fast TAG chart parser with polynomial time complexity. This is surprising given that the XDK constraint parser has exponential time complexity in the worst case.

For our experiment, we applied the most recent version of the XTAG grammar from February 2001, which has a full form lexicon of 45,171 words and 1,230 elementary treyes. The average lexical ambiguity is 28 elementary trees per word, and the maximum lexical ambiguity 360 (for *get*). Verbs are typically assigned more than 100 elementary trees. We developed an encoding of the XTAG grammar into XDG based on ideas from [12] and our encoding of regular dependency grammars (Sect. 7), and implemented these ideas in the XDK.

We tested the XDK with this grammar on a subset of Section 23 of the Penn Treebank, where we manually replaced words not in the XTAG lexicon by appropriate words from the XTAG lexicon. We compared our results with the official XTAG parser: the LEM parser [44], a chart parser implementation with polynomial complexity. For the LEM parser, we measured the time required for building up the

chart, and for the XDK parser, the time required for the first solution and the first 1,000 solutions. Contrary to the LEM parser, the XDK parser does not build up a chart representation for the efficient enumeration of parses. Hence one of the most interesting questions was how long the XDK parser would take to find not only the first but the first 1,000 parses.

We did not use the supertagger included in the LEM package, which significantly increases its efficiency at the cost of accuracy [44]. We must also note that longer sentences are assigned up to millions of parses by the XTAG grammar, making it unlikely that the first 1,000 parses found by the constraint parser also include the *best* parses. This could be remedied with sophisticated search techniques for constraint parsing [14, 39].

We parsed 596 sentences of section 23 of the Penn Treebank whose length ranged from 1 to 30 on an Athlon 64 3000+ processor with 1 GByte of RAM. The average sentence length was 12.36 words. From these 596 sentences, we first removed all those which took longer than a timeout of 30 min. using either the LEM or the XDK parser. The LEM parser exceeded the timeout in 132 cases, and the XDK in 94 cases, where 52 of the timeouts were shared among both parsers. As a result, we had to remove 174 sentences to end up with 422 sentences where neither LEM nor the XDK had exceeded the timeout. They have an average length of 10.73 words.

The results of parsing these remaining 422 sentences is shown in Table 2. Here, the second column shows the time the LEM parser required for building up the chart, and the percentage of exceeded timeouts. The third and fourth column show the times required by the standard XDK parser (using the constraint engine of MOZART/OZ 1.3.2) for finding the first parse and the first 1,000 parses, and the percentage of exceeded timeouts. The fourth and fifth column show the times when replacing the standard MOZART/OZ constraint engine with the new, faster GECODE 2.0.0 constraint library [46], and again the percentage of exceeded timeouts.

Interestingly, despite the polynomial complexity of the LEM parser, the XDK parser not only less often ran into the 30 min timeout, but was also faster than LEM on the remaining sentences. Using the standard MOZART/OZ constraint engine, the XDK found the first parse 3.2 times faster, and using GECODE, 16.8 times faster. Even finding the first 1,000 parses was 1.7 (MOZART/OZ) and 7.8 (GECODE) times faster. The gap between LEM and the XDK parser increased with increased sentence

Table 2 Results of the XTAG parsing experiment

		XDK			
		MOZART/OZ		GECODE	
	LEM	1 parse	1,000 parses	1 parse	1,000 parses
1–30 words	200.47 s	62.96 s	117.29 s	11.90 s	25.72 s
timeouts	132 (22.15%)	93 (15.60%)	94 (15.78%)	60 (10.07%)	60 (10.07%)
1–15 words	166.03 s	60.48 s	113.43 s	11.30 s	24.52 s
timeouts	40 (8.26%)	42 (8.68%)	43 (8.88%)	17 (3.51%)	17 (3.51%)
16–30 words	1204.10 s	135.24 s	229.75 s	29.33 s	60.71 s
timeouts	92 (82.14%)	51 (45.54%)	51 (45.54%)	43 (38.39%)	43 (38.39%)

length. Of the sentences between 16 and 30 words, the LEM parser exceeded the timeout in 82.14% of the cases, compared to 45.54% (MOZART/OZ) and 38.39% (GECODE). Finding the first parse of the sentences between 16 and 30 words was 8.9 times faster using MOZART/OZ, and 41.1 times faster using GECODE. The XDK parser also found the first 1000 parses of the longer sentences faster than LEM: 5.2 times faster using MOZART/OZ and 19.8 times faster using GECODE.

9 Conclusion

The goals of the research reported in this chapter were to classify dependency grammars in terms of their generative capacity and parsing complexity, and to explore their expressive power in the context of a practical system. To reach the first goal, we have developed the framework of *regular dependency grammars*, which provides a link between dependency structures on the one hand, and mildly context-sensitive grammar formalisms such as TAG on the other. To reach the second goal, we have designed a new meta grammar formalism, XDG, implemented a grammar development environment for it, and used this to give novel accounts of linguistic phenomena such as word order variation, and to develop a powerful syntax–semantics interface. Taken together, our research has provided fundamental insights into both the theoretical and the practical aspects of dependency grammars, and a more accurate picture of their usability.

References

1. 45th Annual Meeting of the Association for Computational Linguistics (ACL) (2007).
2. Bader, R., Foeldesi, C., Pfeiffer, U., Steigner, J. Modellierung grammatischer Phänomene der deutschen Sprache mit Topologischer Dependenzgrammatik. Germany: Softwareprojekt, Saarland University (2004).
3. Becker, T., Rambow, O., Niv, M. The derivational generative power, or, scrambling is beyond LCFRS. Tech. Rep., University of Pennsylvania, Philadelphia (1992).
4. Bodirsky, M., Kuhlmann, M., Möhl, M. Well-nested drawings as models of syntactic structure. In: Tenth Conference on Formal Grammar and Ninth Meeting on Mathematics of Language. Edinburgh, UK (2005).
5. Culotta, A., Sorensen, J. Dependency tree kernels for relation extraction. In: 42nd Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 423–429). Barcelona, Spain (2004). DOI 10.3115/1218955.1219009
6. Debusmann, R. A declarative grammar formalism for dependency grammar. Diploma thesis, Saarland University (2001). Http://www.ps.uni-sb.de/Papers/abstracts/da.html
7. Debusmann, R. Multiword expressions as dependency subgraphs. In: Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing. Barcelona/ES (2004).
8. Debusmann, R. Extensible dependency grammar: A modular grammar formalism based on multigraph description. Ph.D. thesis, Universität des Saarlandes (2006).
9. Debusmann, R. The complexity of First-Order Extensible Dependency Grammar. Tech. Rep., Saarland University (2007).
10. Debusmann, R. Scrambling as the intersection of relaxed context-free grammars in a model-theoretic grammar formalism. In: ESSLLI 2007 Workshop Model Theoretic Syntax at 10. Dublin/IE (2007).

11. Debusmann, R., Duchier, D., Koller, A., Kuhlmann, M., Smolka, G., Thater, S. A relational syntax-semantics interface based on dependency grammar. In: Proceedings of COLING 2004. Geneva/CH (2004).
12. Debusmann, R., Duchier, D., Kuhlmann, M., Thater, S. TAG as dependency grammar. In: Proceedings of TAG+7. Vancouver/CA (2004).
13. Debusmann, R., Duchier, D., Niehren, J. The XDG grammar development kit. In: Proceedings of the MOZ04 Conference, Lecture Notes in Computer Science (vol. 3389, pp. 190–201). Springer, Charleroi/BE (2004).
14. Dienes, P., Koller, A., Kuhlmann, M. Statistical A* dependency parsing. In: Prospects and Advances in the Syntax/Semantics Interface. Nancy/FR (2003).
15. Duchier, D., Debusmann, R. Topological dependency trees: A constraint-based account of linear precedence. In: Proceedings of ACL 2001. Toulouse/FR (2001).
16. Egg, M., Koller, A., Niehren, J. The constraint language for lambda structures. Journal of Logic, Language, and Information (2001).
17. Eisner, J., Satta, G. Efficient parsing for bilexical context-free grammars and Head Automaton Grammars. In: 37th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 457–464). College Park, MD, USA (1999). DOI 10.3115/1034678.1034748
18. Gaifman, H. Dependency systems and phrase-structure systems. Information and Control, 8:304–337 (1965).
19. Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M. Prague Dependency Treebank 2.0. Linguistic Data Consortium, 2006T01 (2006).
20. Havelka, J. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL) [1] (pp. 608–615) (2007). URL http://www.aclweb.org/anthology/P/P07/P07-1077.pdf
21. Hays, D.G. Dependency theory: A formalism and some observations. Language, 40(4):511–525 (1964). DOI 10.2307/411934
22. Holan, T., Kuboň, V., Oliva, K., Plátek, M. Two useful measures of word order complexity. In: Workshop on Processing of Dependency-Based Grammars (pp. 21–29). Montréal, Canada (1998).
23. Hotz, G., Pitsch, G. On parsing coupled-context-free languages. Theoretical Computer Science, 161(1–2):205–233 (1996). DOI 10.1016/0304-3975(95)00114-X
24. Hudson, R.A. English Word Grammar. Oxford/UK: B. Blackwell (1990).
25. Huybregts, R. The weak inadequacy of context-free phrase structure grammars. In G. de Haan, M. Trommelen, W. Zonneveld (Eds.), Van periferie naar kern (pp. 81–99). Dordrecht, The Netherlands: Foris (1984).
26. Joshi, A.K. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In: Natural Language Parsing (pp. 206–250). Cambridge: Cambridge University Press (1985).
27. Joshi, A.K., Schabes, Y. Tree-Adjoining Grammars. In G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages (vol. 3, pp. 69–123). New York: Springer (1997).
28. Koller, A., Striegnitz, K. Generation as dependency parsing. In: Proceedings of ACL 2002. Philadelphia/US (2002).
29. Kruijff, G.J.M. Dependency grammar. In: Encyclopedia of Language and Linguistics (2nd edn., pp. 444–450). Amsterdam: Elsevier (2005).
30. Kuhlmann, M. Dependency structures and lexicalized grammars. Doctoral dissertation, Saarland University, Saarbrücken, Germany (2007).
31. Kuhlmann, M., Möhl, M. Mildly context-sensitive dependency languages. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL) [1] (pp. 160–167). URL http://www.aclweb.org/anthology/P07-1021
32. Kuhlmann, M., Möhl, M. The string-generative capacity of regular dependency languages. In: Twelfth Conference on Formal Grammar. Dublin, Ireland (2007).
33. Kuhlmann, M., Nivre, J. Mildly non-projective dependency structures. In: 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Asso-

ciation for Computational Linguistics (COLING-ACL), Main Conference Poster Sessions (pp. 507–514). Sydney, Australia (2006). URL http://www.aclweb.org/anthology/P06-2000
34. Marcus, S. Algebraic Linguistics: Analytical Models, Mathematics in Science and Engineering (vol. 29). New York, USA: Academic Press (1967).
35. McDonald, R., Satta, G. On the complexity of non-projective data-driven dependency parsing. In: Tenth International Conference on Parsing Technologies (IWPT)(pp. 121–132). Prague, Czech Republic (2007). URL http://www.aclweb.org/anthology/W/W07/W07-2216
36. Mel'čuk, I. Dependency Syntax: Theory and Practice. Albany/US: State University Press of New York (1988).
37. Mezei, J.E., Wright, J.B. Algebraic automata and context-free sets. Information and Control, 11(1–2):3–29 (1967). DOI 10.1016/S0019-9958(67)90353-1
38. Mozart Consortium. The Mozart-Oz website (2007). Http://www.mozart-oz.org/
39. Narendranath, R. Evaluation of the Stochastic Extension of a Constraint-Based Dependency Parser. Bàchelorarbeit, Saarland University, Germany (2004).
40. Neuhaus, P., Bröker, N. The complexity of recognition of linguistically adequate dependency grammars. In: 35th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 337–343). Madrid, Spain (1997). DOI 10.3115/979617.979660
41. Nivre, J. Constraints on non-projective dependency parsing. In: Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL) (pp. 73–80). Trento, Italy (2006).
42. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D. The CoNLL 2007 shared task on dependency parsing. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 915–932). Prague, Czech Republic (2007). URL http://www.aclweb.org/anthology/D/D07/D07-1096
43. Quirk, C., Menezes, A., Cherry, C. Dependency treelet translation: Syntactically informed phrasal SMT. In: 43rd Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 271–279). Ann Arbor, USA (2005). DOI 10.3115/1219840.1219874
44. Sarkar, A. Combining supertagging with lexicalized tree-adjoining grammar parsing. Complexity of Lexical Descriptions and its Relevance to Natural Language Processing: A Supertagging Approach. Cambridge, MA: MIT Press (2007).
45. Schulte, C. Programming Constraint Services, Lecture Notes in Artificial Intelligence (vol. 2302). Berlin: Springer-Verlag (2002).
46. Schulte, C., Lagerkvist, M., Tack, G. GECODE—Generic Constraint Development Environment (2007). Http://www.gecode.org/
47. Setz, J. A principle compiler for Extensible Dependency Grammar. Tech. Rep., Bachelorarbeit, Saarland University, Germany (2007).
48. Sgall, P., Hajicova, E., Panevova, J. The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Dordrecht/NL: D. Reidel (1986).
49. Shieber, S.M. Evidence against the context-freeness of natural language. Linguistics and Philosophy, 8(3):333–343 (1985). DOI 10.1007/BF00630917
50. Smolka, G. The Oz programming model. In: J. van Leeuwen (Ed.), Computer Science Today, Lecture Notes in Computer Science (vol. 1000, pp. 324–343). Berlin/DE: Springer-Verlag (1995).
51. Tesnière, L. Éléments de Syntaxe Structurale. Paris, France: Klinksieck (1959).
52. Veselá, K., Havelka, J., Hajičová, E. Condition of projectivity in the underlying dependency structures. In: 20th International Conference on Computational Linguistics (COLING) (pp. 289–295). Geneva, Switzerland (2004). DOI 10.3115/1220355.1220397
53. Vijay-Shanker, K., Weir, D.J., Joshi, A.K. Characterizing structural descriptions produced by various grammatical formalisms. In: 25th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 104–111). Stanford, CA, USA (1987). DOI 10.3115/981175.981190

54. Weir, D.J. Characterizing mildly context-sensitive grammar formalisms. Ph.D. thesis, University of Pennsylvania, Philadelphia, USA (1988). URL http://wwwlib.umi.com/dissertations/fullcit/8908403
55. XTAG Research Group. A Lexicalized Tree Adjoining Grammar for English. Tech. Rep. IRCS-01-03, IRCS, University of Pennsylvania, Philadelphia (2001).
56. Yli-Jyrä, A. Multiplanarity – a model for dependency structures in treebanks. In: Second Workshop on Treebanks and Linguistic Theories (TLT) (pp. 189–200). Växjö, Sweden (2003).

ΩMEGA: Resource-Adaptive Processes in an Automated Reasoning System

Serge Autexier, Christoph Benzmüller, Dominik Dietrich, and Jörg Siekmann

1 Motivation and Historical Background

The ΩMEGA project and its predecessor, the MKRP-system, grew out of the principal dissatisfaction with the methodology and lack of success of the search-based "logic engines" of the 1960s and 1970s.

In the mid-1970s, the paradigm shift from purely search-based general-purpose mechanisms in AI – such as the general problem solver (GPS) to knowledge-based systems as hallmarked by the work of Carl Hewitts [45] – did not leave the theorem proving communities without effect either. Examples are Pat Hayes' article "An Arraignment of Theorem-Proving, or The Logician's Folly" [44] or the well-known quotation that motivated Woody Bledsoe's work:

> Automated Theorem Proving is not the beautiful process we know as mathematics. This is 'cover your eyes with blinders and hunt through a cornfield for a diamond-shaped grain of corn…' Mathematicians have given us a great deal of direction over the last two or three millennia. Let us pay attention to it. (W. Bledsoe, 1986)

These are witnesses to the soul-searching debate about future directions in automated reasoning during that time. These arguments had a profound effect on us in Germany as well, when we started the MKRP initiative within the newly founded basic research institution SFB 813[1] (called Sonderforschungsbereich in German) in 1978, that lasted – due to the long-term perspective of the DFG funding policy in these institutions – for more than a decade. The premise, that no logical system based just on search would ever be able to achieve the mathematical competence of a (human) mathematician, dominated this research from the start. However, classical theorem proving systems based on resolution, matrix or tableaux and relying on highly sophisticated search techniques still witnessed dramatic improvements in

S. Autexier (✉)
DFKI GmbH and Saarland University, 66123 Saarbrücken, Germany
e-mail: autexier@dfki.de

[1] SFB 813 was the predecessor to SFB 378, which is the funding source covered by this chapter. SFB 813 was an important hallmark in Germany as it started AI properly within academia and had profound influence on German Computer Science in general.

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_17,

their performance almost by the day. Thus the overall design of the MKRP system aimed at an approach, which would take these two aspects into account: a strong deduction engine (in this case based on Kowalski's connection graphs) was to be guided by the "supervisor components" that incorporated mathematical knowledge and the myriad of special purpose proof techniques a mathematician has at his or her disposal.

But after almost 10 years of experimentation and system improvements supported by a financial budget that by far exceeded anything that had been granted for automated theorem proving developments at the time, where at times more than a dozen RAs worked on the design, implementation and improvement of the system, we gave up and declared failure.

As Dieter Hutter and Werner Stephan put it ironically in Jörg Siekmann's Festschrift [48]:

> Why did the MKRP-effort fail? Well, it was certainly not a complete failure, but then: why did it not live up to its expectations? After all, it was based on mainstream research assumptions of artificial intelligence, i.e. transporting the water of knowledge based techniques into the intellectual desert of search based automated theorem proving. In Jörg's opinion it is not knowledge-based AI that failed, but their own apparent lack of radicalism. While on the bottom of MKRP there was a graph based, first order theorem proving mechanism that was optimized for a non-informed search, there was the plan of a supervisor module incorporating the necessary domain knowledge in mathematics and controlling effectively the logic engine. But the distance between this general mathematical knowledge to be represented in the supervisor and the low level of abstraction of the logic engine was just too much and the supervisor module never went into existence.

So what went wrong? The research communities working on automated deduction in the nineteen seventies where not exactly known for their lack of confidence and self-esteem: the conference series CADE not only represented one of the most creative and technically mature research communities in AI but automated deduction was without question considered a key component of almost any AI system.

Now, by analogy, anybody in the field would acknowledge that building a car requires more than just building the combustion engine, nevertheless as this is doubtless its key component, whose invention and improvement would need to attract the most outstanding minds of a generation, they better concentrate on this topic first. Replacing "car" by "mathematical assistant system" and looking now back in hindsight it is not apparent whether the "logic engines" of the time (and until today) actually drive the car or more likely just its windscreen wiper: purely search-based logic systems are an important and nice-to-have component, but they certainly do not propel the car. This became common insight in the 1980s except to hardcore combatants, who – fortunately and nevertheless – still increased dramatically the performance of the classical systems, as witnessed inter alia by the yearly TPTP-contest.

So why then did the MKRP approach fail? *"Nothing can be explained to a stone"* is a quotation from John McCarthy that captures it nicely: in order to take advice, a system must have some basic cognitive structures – inborn in biological beings – such that the actual advice given makes sense.

And this is the main flaw with search-based logic engines that work by refutation and normal form representations: the advice given by a mathematical supervisor just does not make sense at this level of "logical machine code". Not only is the distance between the logical representation in the search engine and the level of abstraction of (human) mathematical knowledge representation a long way off, which could possibly be bridged by intermediate levels of representation and abstraction, but the basic mode of operation is just as fundamentally different as, say, flying a plane or the fly of a bird. And unfortunately for this classical research paradigm in AI: the bird (the mathematician) is infinitely better at flying (theorem proving) than its technical counterpart the plane (the deduction engine).

In other words, the supervisor never quite managed to influence the connection graph based deduction engine in any significant way, and although the overall system performance improved steadily in tune with the international development of the field, the grand hopes[2] of 1976 when it all got under way, never materialised.

At times MKRP was by far the strongest system on the international market, but soon the search-based systems would catch up and the pendulum for the prize of the best system used to swing sometimes to our side of the Atlantic but in little time back to our strongest contender and friend Larry Wos and his systems on the American side of the Ocean. In fact this race completely dominated almost a decade of our lives, with Larry Wos calling – sometimes in the middle of the night, because of the time shift: "Hey, we have solved another problem, can you do it as well?"[3]

The general outcome of this race can still best be described by a quote from the summary of the IJCAI-paper [36] in 1980:

> At present the [MKRP-] system performs substantially better than most other automatic theorem proving systems [...] (However) this statement is less comforting than it appears: the comparison is based on measures of the search space and it *totally neglects* the (enormous) resources needed in order to achieve the behaviour described. Within this frame of reference it would be easy to design the 'perfect' proof procedure: the supervisor and the look-ahead heuristics would find the proof and then guide the system without any unnecessary steps through the search space.
>
> Doubtlessly, the TP systems of the future will have to be evaluated in totally different terms, which take into account the *total* (time and space) resources needed in order to find the proof of a given theorem.
>
> But then, is the complex system A, which 'wastes' enormous resources even on relatively easy theorems but is capable of proving difficult theorems, worse than the smart system B, which efficiently strives for a proof but is unable to contemplate anything above the current average TP community? But the fact that system A proves a theorem of which system B is incapable is no measure of performance either, unless there is an objective measure of 'difficulty' (system A may e.g. be tuned to that particular example). If now the difficulty of a theorem is expressed in terms of the resources needed in order to prove it

[2] As expressed in our first research proposal: "MKRP would not only show the usual steady improvements in strength and performance, but a principal leap in performance by some orders of magnitude"

[3] These phone calls constituted a great honour: as opposed to the dozens of experimental systems on the market and presented at CADE, you were now considered a serious contender within the inner circle that distinguished the boys from the men.

> the circulus virtuosus is closed and it becomes apparent that the 'objective' comparison of systems will be marred by the same kind of problems that have marked the attempts to 'objectify' and 'quantify' human intelligence: they measure certain aspects but ignore others.
>
> In summary, although there are good fundamental arguments supporting the hypothesis that the future of ATP research is with the finely tuned knowledge engineered systems as proposed here, there is at present no evidence that a traditional TP with its capacity to quickly generate many ten thousands of clauses is not just as capable.

In a panel discussion with Jörg Siekmann and his friend Larry Wos, Larry put the whole debate nicely into the following joke: "Jörg, why don't you use our smart system OTTER as the supervisor? It will find even difficult proofs and then you use this 'knowledge' to guide your knowledge based system beautifully through the search space without any additional search."

So in the mideighties – although there was still incremental progress – we had to face the question: "Suppose we are given another 10 years of funding, will the system (or any other search-based logic system) ever reach human level performance?"

And as remarked above, our final answer was "NO!". Using the last few years of funding of the SFB 813, we started to look for alternatives: if the "logic engine" can't drive the car, what else could? Woody Bledsoe's succession of systems that incorporated mathematical knowledge surely pointed the way, but they were lacking in generality: if every new piece of mathematics required fresh (LISP) coding, how could we ever obtain a uniform system?[4]

And these weekly discussions, headed by Manfred Kerber, who was at the helm of our theorem proving group at the time, provided the fertile grounds to receive the new seeds coming from Edinburgh with Alan Bundy's ideas on proof planning [24, 54], which were later implemented in their system [27].

1.1 The ΩMEGA *Initiative*

Declaring failure is an honourable retreat in established sciences such as mathematics or physics,[5] with high-risk research projects. Given the general climate of funding for Computer Science in Germany with its stiff competition it came like a miracle that nevertheless we were given another chance within the newly established *Sonderforschungsbereich* SFB 378: the ΩMEGA project was again funded during the entire lifetime (12 years) of the new SFB.

---

[4] Of course a mathematicians brain – by all accounts – is not a "uniform system" either, with its 10^{10} neurons and hundreds of thousands connections each, but building another system for every new piece of mathematics is surely not very satisfying.

[5] Imagine the following: "Well, Mr. President, we promised to send a man to the moon, it has absorbed a bit more than a billion dollars, but – alas – currently our technology is not capable of achieving this."

Embracing Alan Bundy's ideas and further developing them on the way, the proof planner is now the central component among many other ingredients of the ΩMEGA system. It can be seen as the "engine" that drives the search for a proof at a human-oriented level of representation, but – as is to be expected – many more components are required to build a system of potential use for a working mathematician.

Hence the research objective of the ΩMEGA project has been to lay the foundation for these complex, heterogenous and well-integrated assistance systems for mathematics, which support the wide range of typical research, publication and knowledge management activities of a working mathematician. Examples for these activities are computing (for instance algebraic and numeric problems), proving (lemmas or theorems), solving (for instance equations), modelling (by axiomatic definitions), verifying (a proof), structuring (for instance the new theory and knowledge base), maintaining (the knowledge base), searching (in a very large mathematical knowledge base), inventing (your new theorems) and finally writing the paper explaining and illustrating the material in natural language and diagrams. Clearly, some of these activities require a high amount of human ingenuity while others do not and they are thus open to computer support with current AI and Computer Science technology.

Our research is based in particular on the combination of techniques from several subfields of AI including knowledge representation and reasoning, cognitive architectures and multi-agent systems, human computer interaction and user interfaces, as well as machine learning, intelligent tutor systems and finally dialog systems with natural language processing capabilities.

The technical notion for the integration of theses techniques is that of a *resource* and the adaptation of the system to a wide range of resources. In fact, a mathematical assistant system can be considered as a performance enhancing *resource for human users* as well as a *resource for the other software systems* using it. Furthermore, a user friendly mathematical assistance system has to solve a given task within *limited time and space resources*. While executing a task, let us say, automatically planning a proof for a theorem, the system's performance may significantly depend on further *knowledge resources* such as proof methods, theorems and lemmas. Furthermore, the assistant system may exploit *specialised computing and reasoning resources*, for example, an external computer algebra system, a classical automated deduction system or a model generator.

Considering a mathematical assistance system itself as a resource requires the development of different interfaces – for a human user or for other software systems. This in turn poses the problem how the system adapts itself to such conceptually very different means of interaction.

This chapter is organised as follows: Sect. 2 presents proof representation and proof search techniques that utilise knowledge and specialised computing resources. We discuss the representation, authoring, access to and maintenance of knowledge resources in Sect. 3 and specialised computing resources in Sect. 4. Section 5 develops the infrastructures and internal architecture that enables the assistance system to adapt to the different means of interaction.

2 Resource-Adaptive Proof Search

This section presents proof search techniques that exploit different resources to prove a given conjecture. The proof procedures all work on a central, elaborate *proof object*, which supports the simultaneous representation of the proof at different levels of granularity and records also alternative proof attempts.

2.1 Human-Oriented High-Level Proofs

The central component of our computer-based proof construction in ΩMEGA is the TASKLAYER (see Fig. 1). It is based on the CORE-calculus [2] that supports proof development directly at the *assertion level* [81], where proof steps are justified not only by basic logic inference rules but also by definitions, axioms, theorems or hypotheses (collectively called *assertions*).

Subgoals to be shown are stored within the TASKLAYER as *tasks*, which are represented as Gentzen-style multi-conclusion sequents [43]. In addition there are means to define multiple foci of attention on subformulas that are maintained within the actual proof. Each task is reduced to a possibly empty set of subtasks by one of the following proof construction steps: (1) the introduction of a proof sketch [86], (2) deep structural rules for weakening and decomposition of subformulas, (3) the application of a lemma that can be postulated on the fly (and proved later),

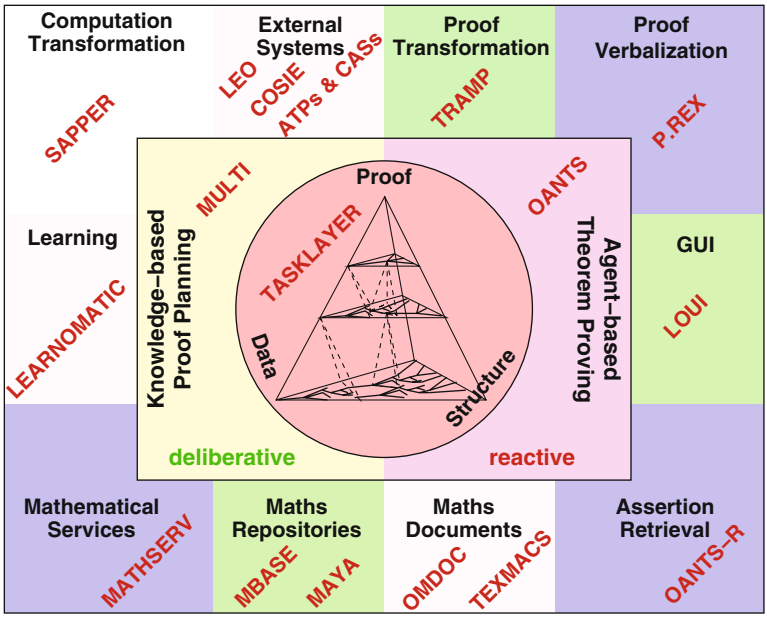


Fig. 1 ΩMEGA overview

(4) the substitution into meta-variables and (5) the application of an *inference*. Inferences are the basic reasoning steps of the TASKLAYER, and comprise assertion applications, proof planning methods or calls to external special systems such as a computer-algebra system, an automated deduction system or a numerical calculation package (see [34, 3] for more details about the TASKLAYER).

2.1.1 Inferences

Intuitively, an *inference* is a proof step with multiple premises and conclusions augmented by (1) a possibly empty set of hypotheses for each premise, (2) a set of *application conditions* that must be fulfilled upon inference application, (3) a set of *completion functions*[6] that compute the values of premises and conclusions from values of other premises and conclusions and (4) an *expansion function* that refines the abstract inference step. Each premise and conclusion consists of a unique name and a formula scheme. Note that we employ the term *inference* in its more general psychological meaning. Taken in this sense, an inference may turn out to be invalid actually, in contrast to the formal logical notion of an *inference rule*.

Additional information needed in the application conditions or the completion functions, such as, for instance, the position of a subterm or the instance of some non-boolean meta-variable, can be specified by additional *parameters* to the inference.

Application conditions are predicates on the values of inference variables and completion functions compute values for specific inference variables from values of other inference variables.

An example of an inference is given in Fig. 2. The inference *subst-m* has two premises p_1, p_2 with formula schemes F and $U = V$, respectively, one conclusion c with formula scheme G and one parameter π. It represents the inference that if we are given a formula F with subterm U at position π and the equation $U = V$, then we can infer the formula G which equals F except that U is replaced by V. The completion functions are used to compute the concrete conclusion formula c, given p_1, p_2 and π. They can also be used for the "backward" direction of the

$$\frac{\mathrm{p_1 : F} \qquad \mathrm{p_2 : U = V}}{\mathrm{c : G}}\ \textit{subst-m}(\pi)$$

Appl.Cond.: $(\mathrm{F}|_\pi = \mathrm{U} \wedge \mathrm{G}|_{\pi \leftarrow \mathrm{V}} = \mathrm{F}) \vee (\mathrm{G}|_\pi = \mathrm{U} \wedge \mathrm{F}|_{\pi \leftarrow \mathrm{V}} = \mathrm{G})$
Completions: $\langle \mathrm{c}, \text{compute-subst-m}\ (\mathrm{p_1}, \mathrm{p_2}, \pi)$
$\langle \mathrm{p_1}, \text{compute-subst-m}\ (\mathrm{p_2}, \mathrm{c}, \pi)$
$\langle \pi, \text{compute-pos}\ \ (\mathrm{p_1}, \mathrm{p_2}) \rangle$
$\langle \pi, \text{compute-pos}\ \ (\mathrm{p_2}, \mathrm{c}) \rangle$

Fig. 2 Inference *subst-m*

[6] The completion functions replace the "outline functions" in previous work [73, 82].

inference to compute the formula p_1, given c, p_2 and π, or to compute the position π at which a replacement can be done. Note that there are two completion functions for computing π. Furthermore, for a given formula F for p_1 and equation $U = V$ for p_2, there are in general more than one possible value for π. Therefore, the completion functions actually compute *streams* of values, each value giving rise to a new instance of the inference.

Inferences can also encode the operational behaviour of domain specific assertions. Consider for instance the domain of set theory and the definition of $\subseteq$:

$$\forall U, V.U \subseteq V \Leftrightarrow (\forall x.x \in U \Rightarrow x \in V)$$

That assertion gives rise to two inferences:

$$\frac{\begin{array}{c}[x \in U] \\ \vdots \\ p : x \in V\end{array}}{c : U \subseteq V} \textit{Def-}\subseteq \qquad \textbf{Appl. Cond.: } x \textit{ new for } U \textit{ and } V$$

$$\frac{p_1 : U \subseteq V \quad p_2 : x \in U}{c : x \in V} \textit{Def-}\subseteq \qquad \textbf{Appl. Cond.: } x \textit{ new for } U \textit{ and } V$$

As a result, we obtain proofs where each inference step is justified by a mathematical fact, such as a definition, a theorem or a lemma.

To illustrate the difference between a typical proof step from a textbook and its formal counterpart in natural deduction consider the assertion step that derives $a_1 \in V_1$ from $U_1 \subseteq V_1$ and $a_1 \in U_1$. The corresponding natural deduction proof is

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\forall U, V.\, U \subseteq V \Leftrightarrow \forall x.x \in U \Rightarrow x \in V}{\forall V.\, U_1 \subseteq V \Leftrightarrow \forall x \in U_1 \Rightarrow x \in V}\forall_E}{U_1 \subseteq V_1 \Leftrightarrow \forall x.x \in U_1 \Rightarrow x \in V_1}\forall_E}{U_1 \subseteq V_1 \Rightarrow \forall x.x \in U_1 \Rightarrow x \in V_1}\Leftrightarrow_E \quad U_1 \subseteq V_1}{\forall x.x \in U_1 \Rightarrow x \in V_1}\Rightarrow_E}{a_1 \in U_1 \Rightarrow a_1 \in V_1}\forall_E \quad a_1 \in U_1}{a_1 \in V_1}\Rightarrow_E$$

Even though natural deduction proofs are far better readable than proofs in machine oriented formalisms such as resolution, we see that they are at a level of detail we hardly find in a proof of a typical mathematical textbook or in a research publication. In the example above, a single assertion step corresponds to six steps in the natural deduction calculus.

Similarly, the lemma $\forall U, V, W.(U \subseteq V \wedge V \subseteq W) \Rightarrow U \subseteq W$ stating the transitivity of $\subseteq$ can be represented by the inference:

$$\frac{p_1 : U \subseteq V \quad p_2 : V \subseteq W}{c : U \subseteq W} \textit{Trans-}\subseteq \tag{1}$$

An eminent feature of the TASKLAYER is that inferences can be applied to *sub*-formulas of a given task. Consider the task to prove that if f is an automorphism on some group G, then the f-image of the Kernel of G is a subset of G.

$$A \subseteq B \Rightarrow f(A) \subseteq f(B) \vdash \mathit{AutoMorphism}(f, G) \Rightarrow f(\mathit{Ker}(f, G)) \subseteq G \qquad (2)$$

where we have the additional hypothesis that if two arbitrary sets are in a subset relation, then so are their images under f.

A deep application of the inference *Trans-* $\subseteq$ matching the conclusion with the subformula $f(\mathit{Ker}(f, G)) \subseteq G$ and the first premise with the subformula $f(A) \subseteq f(B)$ reduces the task in one step to

$$A \subseteq B \Rightarrow (f(A) \subseteq f(B)) \vdash \mathit{AutoMorphism}(f, G) \Rightarrow (\mathit{Ker}(f, G) \subseteq B \wedge f(B) \subseteq G)$$

which can be proved immediately using the definitions of *AutoMorphism* and *Ker*. This one-step inference would not be possible unless we can use (1) to match the subformula within the conclusion of the task (2).

2.1.2 Application Direction of an Inference

The problem is to find all possible ways to apply a given inference to some task, i.e. to compute all possible instantiations of an inference. Typically, some of the parameters as well as some of the formal arguments of the inference are already instantiated. The formal arguments and the parameters of an inference will be collectively called the *arguments* of an inference.

The process starts with a partial argument instantiation (PAI) and we have to find values for the non-instantiated arguments of the inference. These arguments take positions as values within the task or they have formulas as values.

Example 1 Consider the inference *subst-m* of Fig. 2 before, which we want to apply to the task T: $2 * 3 = 6 \vdash 2 * 3 < 7$. Then $pai_1 = \langle \emptyset, \{c \mapsto (10)\}, \emptyset \rangle$ is a partial argument instantiation for the inference *subst-m*, where (10) denotes the position of $2 * 3 < 7$ in the task. As no completion function has been invoked so far, pai_1 is *initial*. It is not *complete* as there is not enough information for the completion functions to compute π given only c; thus p_1, p_2 can not be computed yet.

The extension of a partial argument instantiation consists of an assignment of values to arguments of the inference that are not yet instantiated. There are two possible choices: (1) either assign a task position to a formal argument or (2) to assign a term to a formal argument.

The first kind of update involves searching for possible positions in the task while respecting already introduced bindings. The second kind of update involves no search for further instances in the task, as it uses the completion functions to compute the missing values.

Thus we can separate the updating process into two phases: In the first phase we update the positions by searching the task for appropriate instance of the formula

schemes and in the second phase we only use the completion functions to compute the missing arguments. The general idea is to use as much derived knowledge as possible and then decide whether this knowledge is sufficient for the inference to be drawn. A partial argument instantiation *pai* is called complete, if it contains sufficient information to compute all other values of arguments using completion functions.

Example 2 If we add an instantiation for the argument p_2 in pai_1 we obtain $pai_2 = \langle \emptyset, \{p_2 \mapsto (00), c \mapsto (10)\}, \emptyset \rangle$, where (00) denotes the position of the formula $2 * 3 = 6$ in our task and pai_2 is an extension of pai_1. It is complete, as we can invoke the completion functions to first obtain π and then to obtain p_1.

The configuration of a complete partial argument instantiation describes an *application direction* of the inference. All application directions can be determined by analysing the completion functions of an inference. As an example consider inference *Trans-* $\subseteq$ (p. 396): The application of the inference on task (2) instantiates c with $f(Ker(f, G)) \subseteq G$ and p_1 with $f(A) \subseteq f(B)$, which is represented by the partial argument instantiation $pai := \langle \{p_1 \mapsto f(A) \subseteq f(B)\}, \emptyset, \{c \mapsto f(Ker(f, G)) \subseteq G\} \rangle$. The *configuration* of pai_1, that is the premises and conclusions that are instantiated and those which not, classify this rule application as "backward". The same rule with both premises instantiated but not the conclusion is a "forward" rule.

2.1.3 Representation of Proof

The *proof data structure (PDS)* is at the centre of our system (see Fig. 3) and its task is to maintain the current status of the proof search so far and to represent it at different levels of abstraction and granularity. The PDS is based[7] on the following ideas:

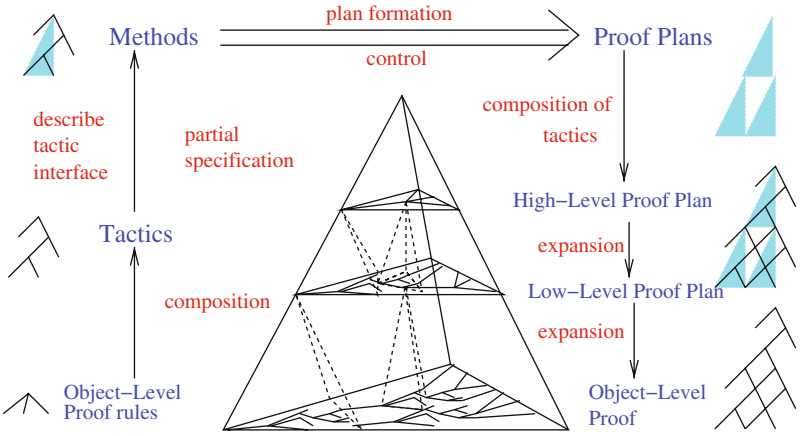


Fig. 3 Proof plan datastructure

[7] It reflects our experience of more than a decade of development of the ΩMEGA system [30, 73–75, 82, 6] as well as ideas from the QUODLIBET system [8].

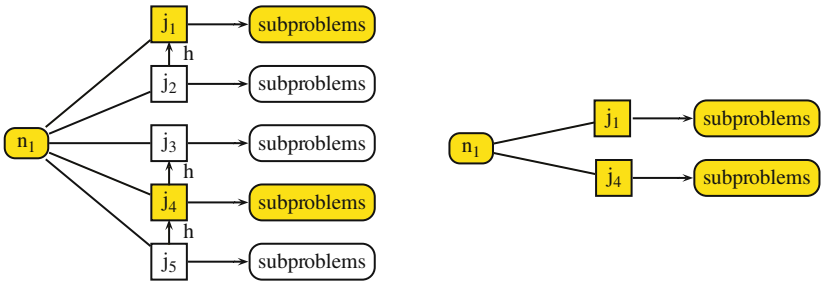
- Each conjectured *lemma* gets its own *proof tree* (actually a directed acyclic graph), whose nodes are (sub-)goals to be shown.
- In this *proof forest*, each lemma can be applied in each proof tree as an inference; either as a lemma in the usual sense or as an induction hypothesis in a possibly mutual induction process, see [87].
- A lemma is a *task* to be proved reductively. A *reduction* step reduces a *goal* to a conjunction of *sub-goals* with respect to a *justification*. These are the proof construction steps of the TASKLAYER.
- Several reduction steps applied to the same goal result in alternative proof attempts, which either represent different proof *ideas* or the same proof idea but at a different level of abstraction or *granularity* (with more or less detail).

The PDS is essentially a directed acyclic graph (dag) whose nodes are labelled with tasks. It has two sorts of links: *justification hyperlinks* represent some relation of a task node to its sub-task nodes, and *hierarchical edges* point from justifications to other justifications which they refine.

This definition allows for alternative justifications and alternative hierarchical edges. In particular, several outgoing justifications of a node n, which are not connected by hierarchical edges, are OR-alternatives. That is, to prove a node n, only the targets of one of these justifications have to be solved. Hence they represent alternative ways to tackle the same problem n. This describes the horizontal structure of a proof. Note further that we can share refinements: for instance, two abstract justifications may be refined by one and the same justification at lower levels.

Hierarchical edges are used to construct the vertical structure of a proof. This mechanism supports both recursive expansion and abstraction of proofs. For instance, in Fig. 4a, the edge from j_2 to j_1 indicates that j_2 refines j_1. The hierarchical edges distinguish between upper layer proof steps and their refinements at a more granular layer.

A proof may be first conceived at a high level of abstraction and then *expanded* to a finer level of granularity. Vice versa, *abstraction* means the process of



(a) PDS-node with all outgoing partially hierarchically ordered justifications, and j_1, j_4 in the set of alternatives. Justifications are depicted as boxes.

(b) PDS-node in the PDS-view obtained for the selected set of alternatives j_1, j_4.

Fig. 4 An example PDS and one of its PDS-views

successively contracting fine-grained proof steps to more abstract proof steps. Furthermore, the PDS generally supports alternative and mutually incomparable refinements of one and the same upper layer proof step. This horizontal structuring mechanism – together with the possibility to represent OR-alternatives at the vertical level – provides very rich and powerful means to represent and maintain the proof attempts during the search for the final proof. In fact, such multidimensional proof attempts may easily become too complex for a human user, but since the user does not have to work simultaneously on different granularities of a proof, elaborate functionalities to access only selected parts of a PDS are helpful. They are required, for instance, for user-oriented presentation of a PDS, in which the user should be able to focus only on those parts of the PDS he is currently working with. At any time, the user can choose to see more details of some proof step or, on the contrary, he may want to see a coarse structure when he is lost in details and can not see the wood for trees.

One such functionality is a *PDS-view* that extracts from a given PDS only a horizontal structure of the represented proof attempts, but with all its OR-alternatives. As an example consider the PDS fragments in Fig. 4.
The node n_1 in the fragment on the left has two alternative proof attempts with different granularities. The fragment on the right gives a PDS-view which results from selection of a certain granularity for each alternative proof attempt, respectively. The set of alternatives may be selected by the user to define the granularity on which he currently wants to inspect the proof. The resulting PDS-view is a slice plane through the hierarchical PDS and is – from a technical point of view – also a PDS, but without hierarchies, that is without hierarchical edges.

2.2 Searching for a Proof

In the following we shall look at our main mechanisms for actually finding a proof and we distinguish two basic modes, knowledge based, i.e. deliberative proof search and reactive proof search.

2.2.1 Knowledge-Based Proof Search

ΩMEGA's main focus is on knowledge-based proof planning [24, 25, 66, 70], where proofs are not conceived in terms of low-level calculus rules, but at a less detailed granularity, that is at a more abstract level, that highlights the main ideas and deemphasises minor logical or mathematical manipulations of formulas. The motivation is to reduce the combinatorial explosion of the search space in classical automated theorem proving by providing means for a more global search control. Indeed, the search space in proof planning tends to be orders of magnitude smaller than at the level of calculus rules [26, 66]. Furthermore, a purely logical proof more often than not obscures its main mathematical ideas.

Knowledge-based proof planning is a paradigm in automated theorem proving, which swings the motivational pendulum back to the AI origins in that it employs

and further develops many AI principles and techniques such as hierarchical planning, knowledge representation in frames and control rules, constraint solving, tactical theorem proving, and meta-level reasoning.

It differs from traditional search-based techniques not least in its level of granularity: The proof of a theorem is planned at an abstract level where an *outline* of the proof is found first. This outline, that is, the abstract proof plan, can be recursively expanded to construct a proof within a logical calculus provided the expansion of the proof plan does not fail.

The building blocks of a proof plan are the plan operators, called *methods* which represent mathematical techniques familiar to a working mathematician. Another important feature is the separation of the knowledge of when to apply a certain technique from the technique itself, which is explicitly stored in *control rules*. Control rules can not only reason about the current goals and assumptions, but also about the whole proof attempt so far.

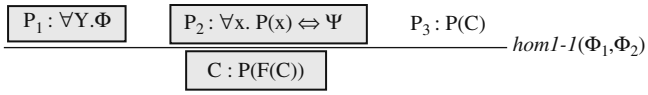
Methods and control rules can employ external systems (for instance, a method may call one of the computer algebra systems) and make use of the knowledge in these systems. ΩMEGA's multi-strategy proof planner MULTI [65, 62, 70] searches for a plan using the acquired methods and strategies guided by the control knowledge in the control rules. In general, proof planning provides a natural basis for the integration of computational systems for both guiding the automated proof construction and performing proof steps.

Knowledge-based proof planning was successfully applied in many mathematical domains, including the domain of limit theorems [66], which was proposed by Woody Bledsoe [22] as a challenge to automated reasoning systems. The general-purpose planner makes use of the mathematical domain knowledge for ε-δ-proofs and of the guidance provided by declaratively represented control rules, which correspond to mathematical intuition about how to prove a theorem in a given situation. Knowledge-based proof planning has also been applied to residue-classes problems where a small set of methods and strategies where sufficient to prove more than 10,000 theorems [64], and to plan "irrationality of $\sqrt[j]{l}$"-conjectures for arbitrary natural numbers j and l [75].

Methods, Control Rules and Strategies

Methods were originally invented by Alan Bundy [24] as tactics augmented with preconditions and effects, called *premises* and *conclusions*, respectively. A method represents a large inference of the conclusion from the premises based on the body of the tactic. The advantage of specifying the effects of a tactic are twofold: (i) the attached tactic need not be executed during the search and (ii) the specification of the tactic may contain additional constraints or control knowledge to restrict the search.

Knowledge-based proof planning expands on these ideas by focusing on encoding domain or problem-specific mathematical methods as proof-planning methods and additionally supports the explicit representation of control knowledge and strategic knowledge. For instance, consider the methods *hom1-1* in Fig. 5.



Appl. Cond.: –
Completions: $\langle \Phi_1 \leftarrow \text{termrploccs}(X, \Phi, C) \rangle$
$\langle \Phi_2 \leftarrow \text{termrploccs}(X, \Phi, F(C)) \rangle$
Expansion: $P_5 : \Phi_2$

Fig. 5 Method *hom1-1*

It encodes a planning operator that uses premises P_1, P_2, P_3 to prove the subgoal C. P_1, P_2 and C are annotated with control information stating in this case that they must be instantiated. Informally, the method describes the following proof situation: If f is a given function, p a defined predicate and the goal is to prove $p(f(c))$, then show $p(c)$ and use this to show $p(f(c))$. Note that P_5 is an open goal that does not occur in the specification and therefore does not directly enter the planning process. The later expansion process will insert p_5 as additional goal in the planning state and then call the planner to close it. To postpone the proof of a goal is an essential feature of methods and provides a means to structure the search space.

Control rules represent mathematical knowledge about how to proceed in the proof planning process. They can influence the planner's behaviour at choice points (such as deciding which goal to tackle next or which method to apply next) by preferring members of the corresponding list of alternatives (for instance, the list of possible goals or the list of possible methods). This way promising search paths are preferred and the search space can be pruned. An example of a control rule is shown in Fig. 6.

```
(control-rule prove-inequality
  (kind methods)
  (IF (and (goal-matches (REL A B))
      (in REL {<,>,<=,>=})))
  (THEN (prefer (TELLCS-B TELLCS-F, ASKCS-B, SIMPLIFY-B,
         SIMPLIFY-F, SOLVE*-B, COMPLEX-ESTIMATE-B,
         FACTORIALESTIMATE-B, SET-FOCUS-B))))
```

Fig. 6 Control rule `prove-inequality`

Its `IF`-part checks whether the current goal is an inequality. If this is the case, it prefers the methods stated after the keyword `prefer` of the rule in the specified order. The general idea of this control rule is that a constraint should be simplified until it can be handled by the constraint solver, which collects constraints in goals and assumptions through the methods `TELLCS-B` and `TELLCS-F`.

Strategies encapsulate fixed sets of methods and control rules and, thus, tackle a problem by some mathematical standard that happens to be typical for this problem. The reasoning as to which strategy to employ for a given problem is an explicit choice point in MULTI. In particular the MULTI system can backtrack from a chosen strategy and commence search with different strategies. An example of a strategy is shown in Fig. 7.

```
(strategy SolveInequality
   (condition inequality-task)
   (algorithm PPlanner)
   (methods COMPLEXESTIMATE-B, TELLCS-B, TELLCS-F, SOLVE*-B ...)
   (crules prove-inequality, eager-instantiage ...)
   (termination no-inequalities)
   )
```

Fig. 7 Strategy SolveInequality

It is applicable for tasks whose formulas are inequalities or whose formulas can be reduced to inequalities. It comprises methods such as `COMPLEXESTIMATE-B` and `TELLCS-B`. It consists of control rules such as `prove-inequality`. The strategy terminates if there are no further tasks containing inequalities.

Detailed discussions of ΩMEGA's method and control rule language can be found in [61, 63]. A detailed introduction to proof planning with multiple strategies is given in [65, 62] and more recently in [70].

2.2.2 Reactive Proof Search

The Ω-ANTS-system was originally developed to support interactive theorem proving [12] in the ΩMEGA-system. Later it was extended to a fully automated theorem prover [14, 78]. The basic idea is to encapsulate each inference into an agent, called an *inference ant*. All ants watch out for their applicability thus generating, in each proof situation, a ranked list of bids for their application. In this process, all inferences are uniformly viewed w.r.t. their arguments, that is, their premises, conclusions as well as other parameters. An inference is applicable if we have found a complete partial argument instantiation and the task of the Ω-ANTS system is to incrementally find complete partial argument instantiations. This starts with an empty partial argument instantiation and searches for instantiations for the missing arguments. This search is performed by separate, concurrent processes (software agents) which compute and report bids for possible instantiations. In other words, each inference ant is realised as a cooperating society of *argument ants*.

The Ω-ANTS Architecture

The architecture consists of two layers. At the bottom layer, bids of possible instantiations of the arguments of individual inferences are computed by argument ants. Each inference gets its own dedicated blackboard and one or several argument ants for each of its arguments. The role of each argument ant is to compute possible instantiations for its designated argument of the inference, and, if successful to record these as bids on the blackboard for this inference. The computation is carried out within the given proof context and by exploiting bids already present on the inference's blackboard, that is, argument instantiations computed by other argument ants working for the same rule. The bids from the bottom layer that are applicable in the current proof state are accumulated at the upper layer and heuristically ranked

by another process. The most promising bid on the upper layer is then applied to the central proof object and the data on the blackboards is cleared for the next round of computations.

Ω-ANTS employs resource-bounded reasoning to guide the search and provides facilities to define and modify the processes at runtime [13]. This enables the controlled integration (for instance, by specifying time-outs) of fully fledged external reasoning systems such as automated theorem provers, computer algebra systems or model generators into the architecture. The use of the external systems is modelled by inferences, usually one for each system. Their corresponding computations are encapsulated into one of the argument ants connected to this inference. For example, consider an inference encapsulating the application of an ATP:

$$\frac{p : \Psi}{c : \Phi}\,\mathrm{ATP}$$
Appl. Cond.: $\Phi = \top$
Completions: $\langle \Psi \leftarrow \mathrm{ATP}(\Phi) \rangle$

The inference contains a completion function that computes the value for its premise given a conclusion argument, that is, an open goal. This completion function is turned into an argument ant for the premise argument. Once an open goal is placed on the blackboard, this argument ant picks it up and applies the prover to it in a concurrent process. While the prover runs, other argument ants for other inferences may run in parallel and try to enable their application. Once the prover found a proof or a partial-proof, it is again written onto the blackboard and subsequently inserted into the proof object if the inference is applied. The semantics of the connections to external reasoners is currently hand-coded, but an ontology could be fruitfully employed as the one suggested in [79].

The advantage of this setup is that it enables proof construction by a collaborative effort of diverse reasoning systems. Moreover, the architecture provides a simple and general mechanism for integrating new reasoners into the system independently of other systems already present. Adding a new reasoning system to the architecture requires only to model it as an inference and to provide the argument ants and the inference ant for it. These ants then communicate with the blackboard by reading and writing subproblems to and from it, as well as writing proofs back to the blackboard in a standardised format.

Communication as a Bottleneck

A main disadvantage of our generic architecture is the communication overhead, since the results from the external systems have to be translated back and forth between their respective syntax and the language of the central proof object. Ω-ANTS has initially been rather inefficient: the case studies reported in [17] show that the larger part of the proof effort is sometimes spent on communication rather than on the actual proof search.

In order to overcome this problem, we devised a new method for the cooperation between two integrated systems via a single inference rule, first presented in [19]. This effectively cuts out the need to communicate via the central proof object.

However, direct bilateral integration of two reasoning systems is difficult if both systems do not share representation formalisms that are sufficiently similar: implementing a dedicated inference for the combination of two particular reasoning systems is more cumbersome than simply integrating each system and its results into Ω-ANTS' central architecture. See [21] for a detailed case study, which evaluates this approach for the cooperation between the higher-order resolution prover LEΩ [11] and the first-order theorem prover VAMPIRE. The general idea is that LEΩ sends the subset of its clauses that do not contain any "real" higher-order literals to VAMPIRE. This has been evaluated with 45 randomly chosen examples about sets, relations and functions: VAMPIRE alone (when using a first-order encoding of the examples) can only solve 34 of these examples while LEΩ in cooperation with VAMPIRE can solve 44 (LEΩ uses a higher-order encoding of the examples). Moreover, LEΩ +VAMPIRE solves these examples more than an order of magnitude faster than VAMPIRE.

3 Knowledge as a Resource

There is a need to organise the different knowledge forms and determine which knowledge is available for which problem. Furthermore, there is a need to avoid redundant information for knowledge maintenance reasons. In the ΩMEGA system we use development graphs [47, 4] as a general mechanism to maintain inferences, strategies and control rules in the system (see Sect. 3.1). In Sect. 3.2 we describe how this knowledge can be formalised and included into the development graph. In Sect. 3.3 we present how inferences can be automatically synthesised from axioms and lemmas maintained in the development graph. Based on inferences, we describe how the knowledge for the planner can be automatically synthesised from inferences (Sect. 3.4) and Sect. 3.5 describes the mechanism to automatically generate a set of argument agents out of the inference descriptions.

3.1 Managing Mathematical Knowledge

The knowledge of the ΩMEGA system is organised in theories that are built up hierarchically by importing knowledge from lower theories via theory morphisms to upper layers. These theories and morphisms are organised respectively as nodes and edges of *development graphs* as implemented in the MAYA system [4] which is the central component of the ΩMEGA system that maintains information about the status of conjectures (unknown, proved, disproved or in-progress) and controls which knowledge is available for which conjecture. MAYA supports the evolution of mathematical theories by a sophisticated mechanism for the *management of change*. Its main feature is to maintain the proofs already done when changing or augmenting the theories.

Each theory in the development graph contains standard information like the signature of its mathematical concepts and their formalisation with axioms, lemmas and theorems. The development graph stores other kinds of knowledge as well, such as specific inferences, strategies, control rules and information that links the symbols with their defining axioms as well as symbol orderings.

Each knowledge item is attached to a specific theory. To make it actually visible in all theories that are built on that theory, the development graph allows to specify how morphisms affect the knowledge item.

For each open lemma in a specific theory, the development graph provides all knowledge items that can potentially be used for the lemma. It is up to the proof procedure to select the relevant parts and possibly transform them into a specific representation for the proof procedure.

3.2 Formalising Mathematical Knowledge

To accommodate a variety of input forms, the ΩMEGA system uses the OMDOC [51] document format as a uniform interface for structured theories.

The OMDOC standard is an XML-language for *O*pen *M*athematical *Doc*uments, which includes structured theories modelled on development graphs and a proof representation formalism that is modelled on ΩMEGA's proof datastructure PDS [5].

Structured theories are not the only knowledge forms we use: Inferences can be automatically synthesised from assertions (Sect. 3.3) and the required planner methods and agents can in turn be synthesised from inferences (Sects. 3.4 and 3.5).

The development and encoding of proof methods by hand is a laborious and therefore we studied automatic learning techniques for this problem in collaboration with colleagues from the LEARNΩMATIC project [49].

When a number of proofs use a similar reasoning pattern, this pattern should be captured by a new method in proof planning and the LEARNΩMATIC system can now learn new methods automatically from a number of well-chosen (positive) examples. Automated learning of proof methods is particularly interesting since theorems and their proofs exist typically in abundance,[8] while the extraction of methods from these examples is a major bottleneck of the proof planning methodology. The evaluation of the LEARNΩMATIC system showed that this approach leads to methods that make the proofs shorter, they reduce search and sometimes they even enable the proof planner to prove theorems that could not be proved before (see [49]).

3.3 From Assertions to Inferences

How can we compute a set of inferences for arbitrary assertions? The intuitive idea is as follows; given the definition of $\subseteq$,

[8] For example, the testbed we developed for proof planning theorems about residue classes consists of more than 70,000 theorems.

$$\forall U, V.U \subseteq V \Leftrightarrow (\forall x.x \in U \Rightarrow x \in V)$$

Reading the equivalence as two implications, this assertion results in the two inferences:

$$\frac{\begin{array}{c}[X \in U] \\ \vdots \\ p : X \in V\end{array}}{c : U \subseteq V} \textit{Def-}\subseteq \qquad \textbf{Appl. Cond.: } X \textit{ new for } U \textit{ and } V$$

$$\frac{p_1 : U \subseteq V \quad p_2 : X \in U}{c : X \in V} \textit{Def-}\subseteq \qquad \textbf{Appl. Cond.: } X \textit{ new for } U \textit{ and } V$$

where U, V and X are meta-variables.

Another example is the definition of the limit of a function

$$\begin{array}{l}\forall f, a, l. \\ \quad \forall \varepsilon.\varepsilon > 0 \Rightarrow \exists \delta.\delta > 0 \Rightarrow \forall x.(0 <\mid x - a \mid \wedge \mid x - a \mid < \delta) \Rightarrow \mid f(x) - l \mid < \varepsilon \\ \qquad \Rightarrow \lim_a f = l\end{array} \tag{3}$$

which can be turned into the inference

$$\frac{\begin{array}{c}[\varepsilon > 0, D > 0, 0 <\mid x - A \mid, \mid x - A \mid < D] \\ \vdots \\ P :\mid F(x) - L \mid < \varepsilon\end{array}}{C : \lim_A F = L} \tag{4}$$

Application Condition: $\mathsf{EV}(\varepsilon, \{F, A, L\}) \wedge \mathsf{EV}(x, \{F, A, L, D\})$
Parameters: ε, x

where F, A, L, D are meta-variables and ε and x are parameters. $\mathsf{EV}(x, \{F, A, L, D\})$ is the application condition requiring that the parameter x should not occur in the instances of $\{F, A, L, D\}$ (i.e. x is an Eigenvariable w.r.t. the instances of F, A, L and D).

The technique to obtain such inferences automatically from assertions [3] follows the introduction and elimination rules of a natural deduction (ND) calculus [43]. Given a formula, in a first phase the ND elimination rules are exhaustively applied to that formula collecting the *Eigenvariable* conditions as we go. This results in a set of inference descriptions with *Eigenvariable* conditions. In a second phase the premises of the inference descriptions are simplified by exhaustively applying ND introduction rules to the premises as well as to possible hypotheses obtained for the premises in that phase. The collected *Eigenvariable* conditions are of the form *"y new w.r.t. S"*, where y is the Eigenvariable and S is a list of constants and meta-variables in which y must not occur (including the symbols in the meta-variable substitutions). Checking these conditions by using the predicate $\mathsf{EV}(y, S)$ we compute inferences of the form

$$\frac{\begin{array}{ccc}[\mathcal{H}_1] & & [\mathcal{H}_n] \\ \vdots & & \vdots \\ P_1 & \ldots & P_n\end{array}}{C}\ \textbf{Parameters}\ (y_1, \ldots, y_n)$$

Application Condition: $\mathsf{EV}(y_1, S_1) \wedge \ldots \wedge \mathsf{EV}(y_m, S_m)$

for every assertion.

3.4 From Inferences to Planner Methods

Inferences are either operational representations of domain axioms, lemmas and theorems or user-defined, domain or problem specific mathematical methods. This may even include specialised computing and reasoning systems. Now, inferences can be applied in many ways (see Sect. 2.1.2), but not all of them contribute to the goal of the current proof plan. Rather, efficient (and controlled) search is only possible if we choose an appropriate subset of the many application directions. For example, suppose the current task is to show $A \subseteq B, x \in B \Rightarrow x \in A \vdash A = B$ and we are given the inference

$$\frac{P_1 : A \subseteq B \quad P_2 : B \subseteq A}{C : A = B}$$

originating from the assertion $A = B \Leftrightarrow A \subseteq B \wedge B \subseteq A$. Suppose further that the proof plan requires to unfold the definitions in the goal first and then to use logical arguments to finish the proof. With respect to the first steps in the proof plan, only those application directions of this inference make sense, in which the conclusion C is instantiated, i.e. the following four partial argument instantiations: $\{P_1, P_2, C\}, \{P_1, C\}, \{P_2, C\}, \{C\}$. A convenient way to select the "right" subset is to specify implicitly the subset $\{ad \in AD | ad \models \Phi\}$ of the application directions AD by a property Φ, such that the application direction ad of interest satisfies it. In other words, Φ determines exactly those application directions which are compatible with the current meta level goal or the current mathematical technique.

In our example, we could specify the subset of interest by requiring that the partial argument instantiations are *backward*, where "backward" means that all conclusions of the inference are instantiated (here C). Another, equally appropriate way would be to characterise the subset of interest as those which reduce the target term with respect to a term ordering.

An inference augmented with the information about the application direction is called a *planning method*. However, given a set of planning methods there is no control information which ranks the inferences, i.e. which controls the choice in case several methods are applicable. This is done by the control rules (see Sect. 2.2.1) that are also maintained in the development graph and provided manually or are part of strategy descriptions.

3.5 From Inferences to Agents

Given an inference, we may wish to automatically synthesise a society of agents as ants in Ω-ANTS in addition to proof planning methods. The problem is that such a society is "fragile" in the sense that removing one agent can result in a non-operational unit that cannot produce useful suggestions or any suggestions at all. Hence we must choose a sufficiently large set of agents such that for each agent there is another agent which produces partial argument instantiations required by the agent.[9] But each agent allocates valuable resources (space and runtime). Thus creating all possible agents would deteriorate the system performance, as it would take too much time to compute any suggestion at all.

Our solution is to generate a so-called *agent creation graph*, whose nodes are equivalence classes on partial argument instantiations, and whose edges are all possible partial argument instantiation updates. A society of agents induces a subgraph of the agent creation graph by restricting the edges corresponding to the society of agents. Reachability of a node from the equivalence class of the empty partial argument instantiation in the induced subgraph means that partial argument instantiations for this equivalence class can be generated by the society of agents. The problem of generating an efficient society of agents such that all equivalence classes are reachable is then a *single-source shortest-path problem*, where we assign positive weights to the edges in the graph. This problem can then be solved by known algorithms [68].

A comparison in [34] of some automatically generated units of argument agents with manually specified argument agents shows that they are almost identical. For details about the algorithm see [34, 3].

4 Specialised Computing and Reasoning Resources

Mathematical theorem proving requires a great variety of skills; hence it is desirable to have several systems with complementary capabilities to orchestrate their use and to integrate their results.

The situation is comparable to, say, a travel booking system that must combine different information sources, such as the search engines, price computation schemes and the travel information in distributed (very) large databases, in order to answer a booking request. The information sources are distributed over the Internet and the access to such specialised travel information sources has to be planned, the results have to be combined and finally there must be a consistency check of the time constraints.

In [89, 90] this methodology was transferred and applied to mathematical problem solving. The MATHSERV system plans the combination of several mathematical

[9] In the old Ω-ANTS approach the agent societies have been carefully specified by the user; here the challenge is to automate this task.

information sources (such as mathematical databases), computer algebra systems (CASs), and reasoning processes such as automated theorem provers (ATPs), constraint solvers (CSs) or model generation systems (MGs).

The MATHSERV system is based on the MATHWEB-SB network of mathematical services [41, 42, 91, 82], which was the first approach for an open and modern software environment that enables modularisation, distribution and networking of mathematical services. This provided the infrastructure for building a society of software agents that render mathematical services by either encapsulating legacy deduction software or other functionalities. The software agents deliver their services via a common mathematical software bus in which a central broker agent provides routing and authentication information. Once the connection to a reasoning system has been established by the broker, the requesting client has to access the reasoning system directly via its API. The software bus and its associated reasoning systems were used not only within the field of automated theorem proving, but also for the semantic analysis of natural language (disambiguating syntactical constraints), verification tasks (proving a verification condition), and others, which resulted sometimes in several thousand theorems per day to be proven routinely for these external users.

The MATHSERV system extends the MATHWEB-SB's client–server architecture by semantic brokering of mathematical services and advanced problem solving capabilities to orchestrate the access to the reasoning systems. The key aspects of the MATHSERV framework are:

Problem-Oriented Interface: a more abstract communication level for MATHWEB-SB, such that general mathematical problem descriptions, can be sent to MATHSERV which in turn returns a solution to that problem. Essentially, we moved from the *service*-oriented interface of MATHWEB-SB to a *problem*-oriented interface for MATHSERV.

Advanced Problem-Solving Capabilities: Typically, a given problem cannot be solved by a single service but only by a combination of several services. In order to support the automatic selection and combination of existing services, the key idea is as follows: an ontology is used for the qualitative description of MATHWEB-SB services and *these descriptions are then used as AI planning operators*, in analogy to the proof planning approach. MATHSERV uses planning techniques [28, 37] to automatically generate a plan that describes how existing services must be combined to solve a given mathematical problem.

We used external systems in the search for a proof in two ways within ΩMEGA: to provide a solution to a subproblem or to give hints for the control of the search. In the first case, the call of a reasoning system is modelled as an inference rule and the output of the incorporated reasoning system is translated and inserted as a subproof into the PDS. This back-translation is necessary for interfacing systems that operate at different levels of granularity, and also for a human-oriented display and inspection of a partial proof. In the other case, where the external system is used

to compute values that may be used to guide the search process, the system can be called by a completion function or from within control rules.

The following external systems were integrated and used in the ΩMEGA system over the years:

Computer Algebra Systems (CAS) provide symbolic computation, which can be used to compute hints to guide the proof search (such as witnesses for existential variables), or, second, to perform some complex algebraic computation such as to normalise or simplify terms. In the latter case the symbolic computation is directly translated into proof steps in ΩMEGA. CASs are integrated via the transformation and translation module SAPPER [77, 50]. Currently, ΩMEGA uses the systems MAPLE [29] and GAP [71].

Automated Deduction Systems (ATP) are used to solve subgoals, currently the first-order provers BLIKSEM [33], EQP [60], OTTER [58], PROTEIN [9], SPASS [85, 42], WALDMEISTER [46, 42], the higher-order systems TPS [1], LEΩ [10, 11] and VAMPIRE [32]. The first-order ATPs were connected via TRAMP [59], which is a proof transformation system that transforms resolution-style proofs into assertion-level ND-proofs which were then integrated into ΩMEGA's PDS. The TPS system generates ND-proofs directly, which could then be further processed and checked with little transformational effort [16].

Model Generators (MG) provide either witnesses for free (existential) variables, or counter-models, which show that some subgoal is not a theorem. ΩMEGA used the model generators SATCHMO [55, 54] and SEM [88].

Constraint Solver (CS) construct mathematical objects with theory-specific properties as witnesses for free (existential) variables. Moreover, a constraint solver can reduce the proof search by checking for inconsistencies of constraints. ΩMEGA employed $\mathcal{CoSIE}$ [69, 50, 92], a constraint solver for inequalities and equations over the field of real numbers.

Automated Theory Formation systems (ATF) explore mathematical theories and search for new properties. The HR system is an ATF system in the spirit of Doug Lenat's AM, which conjectures mathematical theories given empirical data [31]. ΩMEGA used the HR system to provide instances for meta-variables that satisfy some required properties. The MATHSAID system proves and identifies theorems (lemmas, corollaries, etc.) from a given set of axioms and definitions [57]. MATHSAID was used by ΩMEGA to derive interesting lemmas for given mathematical theories which would enable the ATPs to prove theorems they could not prove without these lemmas.

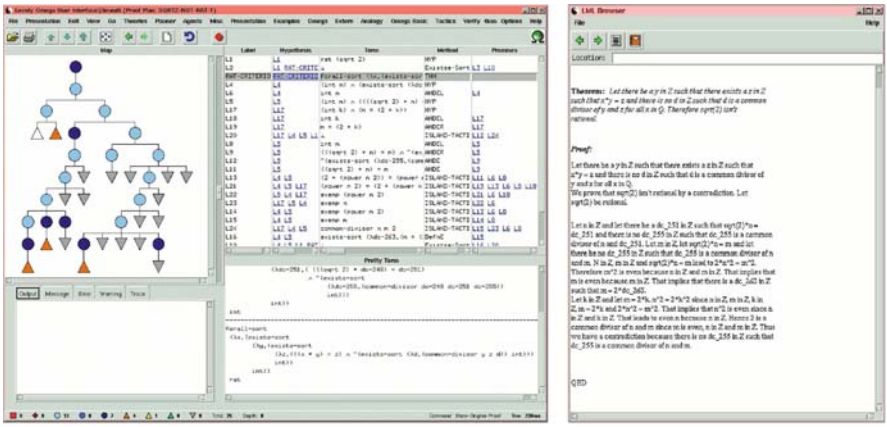
5 Ωmega as an Adaptive Resource

If a mathematical assistance system is to be used as a resource by other systems as well as by users with different skills and backgrounds, we have to redesign the architecture of the system to make it adaptive. We present the research and development

for the interaction with a human-users in Sect. 5.1 and discuss the interaction with other software systems in Sect. 5.2.

5.1 Adaptation to Users with Different Skills

At an early stage of development the OMEGA research group addressed mainly the interaction between the proof assistant system and a human user and for this end we developed an elaborate graphical user interface LΩUI [72] (see Fig. 8a).



(a) The linked proof window views

(b) The natural language presentation of proofs with *P.rex*

Fig. 8 LΩUI: the first graphical user interface of ΩMEGA

The three inter-connected windows present the shape of the central proof tree (left), information about the nodes in the tree (upper right) and the pretty printing of the complete formula of a selected node (lower right). There is also support to switch between different levels of granularity at which a proof can be presented and it is also possible to browse through mathematical theories in an HTML-like viewer. These functionalities were targeted towards a user, who has no knowledge about the actual implementation and the programming language, but who is familiar with the main concepts of formal logic, natural deduction proofs, proof planning methods and tactics.

Also in the early 1980s, members of the group began to research the presentation of proofs in a textbook style form (see Fig. 8b), which does not pre-suppose skills in formal logic from the user. The translation of resolution proofs into natural deduction and the subsequent restructuring techniques for an improved presentation were early results [52, 53]. Based on these developments Xiarong Huang developed the PROVERB system [81], a landmark at its time, which translated these ND-proofs into well-structured natural language texts. Today we use the P.rex-system [38–40], which is based on PROVERB, but presents the proof in a user-adaptive style, i.e. the mode of presentation and abstraction is relative to the skills of the user. The quality

of the proof presentation generated by the P.rex-system is still a corner-stone in the area of proof presentation and the overall development in this field within the last three decades is the subject of the forthcoming textbook [76].

More recent work on human interaction with the system followed a different approach to make it more acceptable to the mathematical community. The mathematical assistance system must be integrated with the software that the users already employ, like standard text processing systems (such as LaTeX) for the preparation of documents. TEXMACS [80] is a scientific text-editor that provides professional typesetting and supports authoring with powerful macro definition facilities like those in LaTeX, but the user works on the final document ("What you see is what you get", WYSIWYG). As a first step we integrated the ΩMEGA system into TEXMACS using the generic mediator PLATΩ [84]. In this setting the formal content of a document is amenable to machine processing, without imposing any restrictions on how the document is structured and which language is used in the document. The PLATΩ system [83] developed in the DFG-project VERIMATHDOC transforms the representation of the formal content of a document into the representation used in a proof assistance system and maintains the consistency between the two representations throughout potential changes.

In turn, the ΩMEGA system provides its support now transparently within the text-editor TEXMACS. Figure 9 shows typical example documents on the screen.

Figure 9a shows how the author can formalise mathematics: Based on the formal representation obtained by PLATΩ, the ΩMEGA system provides its support context-sensitively as a menu inside the text-editor. Figure 9b shows such a menu generated by the system that displays the different assertions which can be applied in the actual proof situation. The proof parts generated by ΩMEGA are patched into the document using natural language patterns. Current work is concerned with adapting the proof presentation techniques as used in *P.rex* to this setting. More details about that integration and the PLATΩ system can be found in [83, 84].

All of this required the following changes in the architecture of the system: First, we need a clean interface with the text-editor TEXMACS. The role of this interface

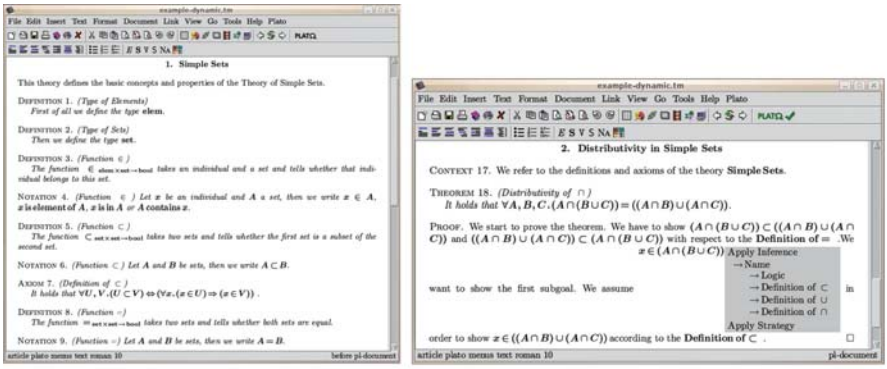


Fig. 9 The user perspective of the support offered by ΩMEGA inside the text-editor TEXMACS via PLATΩ

is to establish and maintain the correspondence of the objects in the document and their counterparts within ΩMEGA. Based on the development graph, the definitions, axioms and theorems in a TEXMACS document are grouped into theories and they have a one-to-one correspondence to the development graph structure. The notion of a "PDS view" (Sect. 2.1.3) is the key prerequisite to consistently link the proof in the document with the respective part of the much more elaborate proof representation PDS in ΩMEGA: Each manually written proof step in the document is modelled as a proof sketch in the PDS, which has to be verified later-on.

Furthermore, the author of a document usually writes many different proofs for different theorems in different theories before the document is finished. This corresponds to multiple, parallel proof attempts in ΩMEGA: the development graph maintains the multiple ongoing proof attempts and determines which assertions in the document are visible for which proof attempt.

Having that infrastructure in place was the key to turn the ΩMEGA system into a server, that can provide mathematical assistance services for multiple documents in parallel sessions.

A second issue is that the author changes his document usually many times. Hence the proof assistance system has to be able to deal efficiently with non-monotonic changes not only inside a theory but also within the proofs. For instance, deleting an axiom from a theory should result in pruning or at least invalidating proof steps in all proofs that relied on that axiom. Furthermore, if by such an action a proof of some theorem is invalidated, then all other proof steps that used this theorem must be flagged and invalidated in turn. The immediate "solution", i.e. to automatically re-execute all proof procedures, is not an option in this setting: it would be too slow and short response times are an issue when the author of the document has to wait when "simply" deleting an axiom. Furthermore, re-executing the proof procedures may generate different proofs: since the proofs within ΩMEGA have to be synchronised with the proofs in the text-editor, this may result in drastic invasive (fully automatic) rearrangements of the document. Such a system behaviour would most certainly jeopardise the acceptance of the system.

For these reasons we integrated a sophisticated and fine granular truth-maintenance system, which tracks all dependencies between elements of a theory and their use in other theories and proofs (see [7] for details).

5.2 Adaptation to Different Software Systems

Just as ΩMEGA is used now as a subsystem within TEXMACS it could be used by other software systems such as a program verification tool or a software development platform. Yet another application area we are currently working on is the integration of ΩMEGA into ACTIVEMATH, an e-learning system for mathematics [67].

More specifically, the DIALOG project [18] studies natural language-based tutorial dialogs when teaching how to prove a theorem. Within a tutorial dialogue, the student has to prove a theorem interactively with the system. The system provides feedback to the student's input, corrects faulty steps and aids the student in finding a

solution, with the overall goal to convey specific concepts and techniques of a given mathematical domain.

Due to the flexible and unpredictable nature of a tutorial dialogue it is necessary to dynamically process and analyse the informal input to the system, including linguistic analysis of the informal input to the system, evaluation of utterances in terms of *soundness*, *granularity* and *relevance*, and *ambiguity resolution* at all levels of processing. ΩMEGA is used to (i) represent the mathematical theory within which the proof exercise is carried out, i.e. the definitions, axioms and theorems of the mathematical domain, (ii) to represent the ongoing proof attempts of the student, in particular the management of ambiguous proof states resulting from underspecified or ambiguous proof attempts, (iii) to maintain the mathematical knowledge the student is allowed to use and to react to changes of this knowledge, and (iv) to reconstruct intermediate steps necessary to verify a step entered by the student, thereby resolving ambiguity and underspecification.

The main problem in such a setting is the high-level and informal nature of human proofs: a classical automated deduction system is of little help here and these developments are only possible now, because of the high-level proof representation and proof planning techniques.

5.2.1 Checking the Correctness

The proof steps entered by the student are statements the system has to analyse with respect to its correctness. A proof step ideally introduces new hypotheses and/or new subgoals along with some justification how they have been obtained. If this information is complete and correct, the verification amounts to a simple check. However, in a tutorial setting, this is not the typical situation. The more likely and, from our point of view, more interesting case is that the statement is incomplete or faulty. Note that an incomplete proof step is not necessarily faulty: when writing proofs, humans typically omit information considered unimportant or trivial. Simply noting that a proof step is false or just incomplete is not a useful hint for the student, so we need a more detailed analysis. If no justification is given, but the hypotheses and subgoals can be correctly derived, the missing justification has to be computed. Conversely, if there is a justification, but the hypothesis or subgoal are missing, the missing parts should be returned by the system. If one of them is false, the system should return a corrected one.

In the sequel we show some typical phenomena extracted from a corpus of tutorial dialogues collected in the Wizard-of-Oz experiments between students and experienced math teachers [20]. Figure 10 shows excerpts from collected dialogues, where the tutor's statements are marked with a capital **T** and the student's utterances with a capital **S**.

Underspecification: The proof step entered by the student is often not fully specified and information may be missing. Utterance **S1** in Fig. 10 is an example of this underspecification which appear throughout the corpus. The proof step in **S1** includes the application of set extensionality, but the rule is not

T: Please prove $(R \circ S)^{-1} = S^{-1} \circ R^{-1}$	**S1a:** we consider the subgoals
S1: let $(x, y) \in (R \circ S)^{-1}$	$(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$
T2: correct	and $(R \circ S)^{-1} \supseteq S^{-1} \circ R^{-1}$
S3: hence $(y, x) \in (S \circ R)$	**S1b:** first, we consider the
T4: incorrect	subgoal $(R \circ S)^{-1} \subseteq S^{-1} \circ R^{-1}$

Fig. 10 *left:* Example dialog between a tutor (**T**) and a student (**S**). *right:* Two alternative ways of how the student started to solve the exercise

stated explicitly. Also the student does not say which of the two subgoals introduced by set extensionality he is now proving, nor does he specify that there is a second subgoal. Further, the number of steps needed to reach this proof state is not given. Part of the task of analysing such steps is to instantiate the missing information so that the formal proof object is complete.

Incomplete Information: In addition to issues of underspecification, there may be crucial information missing for the formal correctness analysis. For instance the utterance **S1** is clearly a contribution to the proof, but since the step only introduces a new variable binding, there is no assertion whose truth value can be checked. However, anticipating that the student wants to use the subset definition $A \subseteq B \Leftrightarrow x \in A \Rightarrow x \in B$ allows us to determine that the new variable binding is useful. Utterance **S1b** is also a correct contribution, but the second subgoal is not stated. This is however necessary in order to establish the equality of the two sets. These examples show that the verification in this scenario is not simply a matter of checking logical correctness.

Ambiguity: Ambiguity pervades all levels of the analysis of natural language and mathematical expressions. Even in fully specified proof steps an element of ambiguity may remain. For example in any proof step which follows **S1a**, we do not know which subgoal the student has decided to work on. Also, when students state formulas without natural language expressions, such as "hence" or "conjecture", it is not clear whether the formula is a newly derived fact or a newly introduced conjecture. Again, this type of ambiguity can only be resolved in the context of the current proof. When no resolution is possible, the ambiguity must be propagated and this must be done by maintaining multiple parallel interpretations, which are retained until enough information is available later on in the proof attempt.

5.2.2 Cognitive Proof States

A well-known phenomenon with underspecified or faulty proof steps is that there is in general more than one reasonable reconstruction. Each reconstruction directly influences the analysis of the subsequent proof step, that is, a subsequent step can be classified to be correct with respect to one reconstruction, but not with respect

to another. Hence, it is necessary to determine and maintain all possible reconstructions, which we call *cognitive proof states*. Ambiguities which cannot immediately be resolved are propagated as parallel cognitive proof states until enough information is available for their resolution.

Technically, the PDS is used to simultaneously represent all of the possible cognitive proof states of the student, each represented by an agenda. Initially, there is only one cognitive proof state, containing the initial task $T = \Gamma \vdash \Delta$ where Γ denotes the assumptions and Δ is the subgoal to be shown.

Updating the Cognitive Proof States

Given a set of possible cognitive proof states and a preprocessed utterance s, all possible successor states have to be determined, which are consistent with the utterance s. Each utterance is possibly – but not necessarily – annotated with information about whether the step represents a new lemma or whether it is supposed to contribute to the overall proof.

For each given cognitive proof state, we determine the successor states that are consistent with the utterance s. If no such successor state can be found this cognitive state is deleted. If several, alternative successor states can be found, i.e. the utterance s is ambiguous, they replace the given cognitive state.

That procedure also resolves ambiguities introduced in previous proof steps by deleting all cognitive states that are no longer consistent with the current utterance s. If all cognitive proof states are deleted, i.e. no successor state is found for any of the given cognitive states, the step is classified as *incorrect*.

The overall result is a confirmation of whether the step could be verified, along with the side-effect that the PDS has been updated to contain exactly the possible cognitive proof states resulting from the performance of the step. More details about the update process are given in [35, 23] and a predecessor system has been described in [15].

6 Future Research

We now want to improve the system quality of ΩMEGA, such that we can train users to author documents with formal logical content. The ΩMEGA system now provides an adequate environment for this endeavour, because its high-level proof representation and proof planning techniques presuppose little knowledge – the main hurdle which typically hampers the use of such systems. Furthermore, ΩMEGA's capabilites to adapt to different users and usages provides a basis for the integration into standard text preparation systems and e-learning environments.

First, we want to support authoring and maintenance of documents with formal logical content such that the author can formulate new concepts, conjectures and proofs in a document. Furthermore, we want to integrate other modalities like diagrams both to describe mathematical content and use it within mathematical proof.

Second, we want to further increase usability of formal reasoning tools by further developing the logical foundations of assertion-level proofs and automate proof search either by proof planning directly on the assertion-level or by transforming proofs obtained from classical automated deduction systems; a key question here is how to characterise and search for "good" proofs. Furthermore, we plan to automate proof search in large, structured theories, where, to date, human guidance of the proof procedures is indispensable, even for theorems that are simple by human standards. We will research how to exploit the structures in large theories not only to search for proofs but also to synthesise new interesting knowledge using automated theory formation [56].

Finally, we want to support the training of students in using formal reasoning tools. Rather than teaching students mathematical proof by forcing them to do a proof in a typical formal calculus, we want to allow the student to freely build any valid proof of the theorem at hand. On the tutoring side, this gives the freedom to adapt the tutoring to the student's skills: less experienced students will be taught more rigid proof styles that come close to the proof style enforced by classical formal calculi, while this is not imposed for more experienced students. In this context, we will further develop domain-independent criteria to dynamically evaluate the correctness, granularity and relevance of user uttered proof steps, provide domain-independent and domain-specific didactic strategies exploiting the dynamic proof step analysis capabilities of the ΩMEGA system, and exploit them to generate useful hints for the student.

Acknowledgments The mathematical assistant system ΩMEGA (and its predecessor MKRP) evolved over a time span of more than 25 years: from its original conception at Karlsruhe in the years 1976 and after, the MKRP system became one of the strongest deduction systems at the time, racing against the succession of systems of Larry Wos and his associates for more than a decade with Christoph Walther at the helm of MKRP and later, when Christoph obtained his professorship, Norbert Eisinger took over. The paradigm shift to knowledge-based proof planning was carried out with Manfred Kerber as project leader to be succeeded by Michael Kohlhase, when Manfred became a lecturer in Britain.

The new ΩMEGA system was developed with Christoph Benzmüller as the last captain at the steering wheel before Serge Autexier now became the current project leader. All in all more than 50 research assistants worked with us on these developments over the time and their contributions are greatly acknowledged.[10]

References

1. Andrews, P., Bishop, M., Issar, S., Nesmith, D., Pfenning, F., Xi, H. TPS: A theorem proving system for classical type theory. Journal of Automated Reasoning 16(3):321–353 (1996).
2. Autexier, S. The CoRe calculus. In R. Nieuwenhuis, (Ed.), Proceedings of the 20th International Conference on Automated Deduction (CADE-20) (vol. 3632). LNAI, Tallinn, Estonia: Springer (2005).

[10] See `http://www.dfki.de/~siekmann` for an explicit account of all PhD students and their publications.

3. Autexier, S., Dietrich, D. Synthesizing proof planning methods and oants agents from mathematical knowledge. In J. Borwein, B. Farmer, (Eds.), Proceedings of MKM'06 (vol. 4108, pp. 94–109). LNAI, London: Springer (2006).
4. Autexier, S., Hutter, D. Formal software development in MAYA. In D. Hutter, W. Stephan, (Eds.), Festschrift in Honor of J. Siekmann (vol. 2605). LNAI, Springer (2005).
5. Autexier, S., Sacerdoti-Coen, C. A formal correspondence between omdoc with alternative proofs and the lambdabarmumutilde-calculus. In J. Borwein, B. Farmer, (Eds.), Proceedings of MKM'06 (vol. 4108, pp. 67–81). LNAI, Springer (2006).
6. Autexier, S., Benzmüller, C., Dietrich, D., Meier, A., Wirth, C.P. A generic modular data structure for proof attempts alternating on ideas and granularity. In M. Kohlhase, (Ed.), Proceedings of the 5th International Conference on Mathematical Knowledge Management (MKM'05) (vol. 3863, pp. 126–142). LNAI, Springer (2006).
7. Autexier, S., Benzmüller, C., Dietrich, D., Wagner, M. Organisation, transformation, and propagation of mathematical knowledge in omega. Journal of Mathematics in Computer Science, 2(2):253–277 (2008).
8. Avenhaus, J., Kühler, U., Schmidt-Samoa, T., Wirth, C.P. How to prove inductive theorems? QUODLIBET! In: Proceeding of the 19th International Conference on Automated Deduction (CADE-19) (pp. 328–333). Springer, no. 2741 in LNAI (2003).
9. Baumgartner, P., Furbach, U. PROTEIN, a PROver with a Theory INterface. In A. Bundy, (Ed.), Proceedings of the 12th Conference on Automated Deduction (pp. 769–773). Springer, no. 814 in LNAI, (1994).
10. Benzmüller, C. Equality and extensionality in higher-order theorem proving. PhD thesis, Department of Computer Science, Saarland University, Saarbrücken, Germany (1999).
11. Benzmüller, C., Kohlhase, M. LEO – a higher-order theorem prover. In C. Kirchner, H. Kirchner, (Eds.), Proceedings of the 15th International Conference on Automated Deduction (CADE-15) (pp. 139–143). Lindau, Germany: Springer, no. 1421 in LNAI (1998).
12. Benzmüller, C., Sorge, V. A blackboard architecture for guiding interactive proofs. In F. Giunchiglia, (Ed.), Proceedings of 8th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'98). Springer, no. 1480 in LNAI (1998).
13. Benzmüller, C., Sorge, V. Critical agents supporting interactive theorem proving. In P. Borahona, J.J. Alferes, (Eds.), Proceedings of the 9th Portuguese Conference on Artificial Intelligence (EPIA'99) (pp. 208–221). Springer, Evora, Portugal, no. 1695 in LNAI (1999).
14. Benzmüller, C., Sorge, V. Ωants – An open approach at combining Interactive and Automated Theorem Proving. In M. Kerber, M. Kohlhase, (Eds.), 8th Symposium on the Integration of Symbolic Computation and Mechanized Reasoning (Calculemus-2000), AK Peters (2000).
15. Benzmüller, C., Vo, Q. Mathematical domain reasoning tasks in natural language tutorial dialog on proofs. In M. Veloso, S. Kambhampati, (Eds.), Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05) (pp. 516–522). Pittsburgh, Pennsylvania, USA: AAAI Press/The MIT Press, (2005).
16. Benzmüller, C., Bishop, M., Sorge, V. Integrating TPS and ΩMEGA. Journal of Universal Computer Science, 5:188–207 (1999).
17. Benzmüller, C., Jamnik, M., Kerber, M., Sorge, V. Agent based mathematical reasoning. Electronic Notes in Theoretical Computer Science, Elsevier, 23(3):21–33 (1999).
18. Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayová, I., Pinkal, M., Siekmann, J., Tsovaltzi, D., Vo, B.Q., Wolska, M. Tutorial dialogs on mathematical proofs. In: Proceedings of IJCAI-03 Workshop on Knowledge Representation and Automated Reasoning for E-Learning Systems (pp. 12–22). Acapulco, Mexico (2003).
19. Benzmüller, C., Sorge, V., Jamnik, M., Kerber, M. Can a higher-order and a first-order theorem prover cooperate? In F. Baader, A. Voronkov, (Eds.), Proceedings of the 11th International Conference on Logic for Programming Artificial Intelligence and Reasoning (LPAR) (pp. 415–431). Springer, no. 3452 in LNAI (2005).
20. Benzmüller, C., Horacek, H., Lesourd, H., Kruijff-Korbayová, I., Schiller, M., Wolska, M. A corpus of tutorial dialogs on theorem proving; the influence of the presentation of the study-material. In: Proceedings of International Conference on Language Resources and Evaluation (LREC 2006). ELDA, Genova, Italy (2006).

21. Benzmüller, C., Sorge, V., Jamnik, M., Kerber, M. Combined reasoning by automated cooperation. Journal of Applied Logic, 6(3):318–342 (2008).
22. Bledsoe, W. Challenge problems in elementary calculus. Journal of Automated Reasoning, 6:341–359 (1990).
23. Buckley, M., Dietrich, D. Integrating task information into the dialogue context for natural language mathematics tutoring. In B. Medlock, D. Ó Séaghdha, (Eds.), Proceedings of the 10th Annual CLUK Research Colloquium, Cambridge, UK (2007).
24. Bundy, A. The use of explicit plans to guide inductive proofs. In E. Lusk, R.A. Overbeek (Ed.), Proceeding of the 9th conference on Automated Deducation no. 310 in LNCS (pp. 111–120). Argonne, Illinois, USA: Springer (1988).
25. Bundy, A. A science of reasoning. In J.-L. Lasser, G. Plotkin (Eds.), Computational Logic: Essays in Honor of Alan Robinson (pp. 178–199). Cambridge, MA: MIT Press (1991).
26. Bundy, A. A critique of proof planning. In: Computational Logic: Logic Programming and Beyond (Kowalski Festschrift) (vol. 2408, pp. 160–177). LNAI, Springer (2002).
27. Bundy, A., van Harmelen, F., Horn, C., Smaill, A. The oyster-clam system. In M.E. Stickel, (Ed.), Proceedings of the 10th Conference on Automated Deduction (vol. 449, pp. 647–648). Springer Verlag, LNAI (1990).
28. Carbonell, J., Blythe, J., Etzioni, O., Gil, Y., Joseph, R., Kahn, D., Knoblock, C., Minton, S., Pérez, M.A., Reilly, S., Veloso, M., Wang, X. PRODIGY 4.0: The Manual and Tutorial. CMU Tech. Rep. CMU-CS-92-150, Carnegie Mellon University (1992).
29. Char, B., Geddes, K., Gonnet, G., Leong, B., Monagan, M., Watt, S. First Leaves: A Tutorial Introduction to Maple V. Springer, New York (1992).
30. Cheikhrouhou, L., Sorge, V. PDS – a three-dimensional data structure for proof plans. In: Proceedings of the International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA'2000) (2000).
31. Colton, S. Automated Theory Formation in Pure Mathematics. Distinguished Dissertations, Springer (2002).
32. de Nivelle, H. Bliksem 1.10 user manual. Tech. Rep., Max-Planck-Institut für Informatik (1999).
33. Dietrich, D. The task-layer of the ΩMEGA system. Diploma thesis, FR 6.2 Informatik, Universität des Saarlandes, Saarbrücken, Germany (2006).
34. Dietrich, D., Buckley, M. Verification of proof steps for tutoring mathematical proofs. In R. Luckin, K.R. Koedinger, J. Greer, (Eds.), Proceedings of the 13th International Conference on Artificial Intelligence in Education (vol. 158, pp. 560–562). Los Angeles, USA: IOS Press (2007).
35. Dijkstra, E.W. A note on two problems in connexion with graphs. Numerische Mathematik, 1:269–271 (1959).
36. Eisinger, N., Siekmann, J., Smolka, G., Unvericht, E., Walther, C. The markgraf karl refutation procedure. In: Proceedings of the Conference of the European Society for Artificial Intelligence and Simulation of Behavior. Amsterdam, Netherlands (1980).
37. Erol, K., Hendler, J., Nau, D. Semantics for hierarchical task network planning. Tech. Rep. CS-TR-3239, UMIACS-TR-94-31, Computer Science Department, University of Maryland (1994).
38. Fiedler, A. Using a cognitive architecture to plan dialogs for the adaptive explanation of proofs. In Dean, T. (Ed.), Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 358–363). Morgan Kaufmann, Stockholm, Sweden (1999).
39. Fiedler, A. Dialog-driven adaptation of explanations of proofs. In B. Nebel, (Ed.), Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1259–1300). Morgan Kaufmann, Seattle, WA (2001).
40. Fiedler, A. User-adaptive proof explanation. PhD thesis, Department of Computer Science, Saarland University, Saarbrücken, Germany (2001).
41. Franke, A., Kohlhase, M. System description: MathWeb, an agent-based communication layer for distributed automated theorem proving. In H. Ganzinger, (Ed.), Proceedings of the 16th conference on Automated Deduction, no. 1632 in LNAI, Springer (1999).

42. Ganzinger, H. (Ed.), Proceedings of the 16th Conference on Automated Deduction, no. 1632 in LNAI, Springer (1999).
43. Gentzen, G. The Collected Papers of Gerhard Gentzen (1934–1938). In: Szabo, M. E (ed), North Holland, Amsterdam (1969).
44. Hayes, P., Anderson, D.B. An arraignment of theorem-proving; or, the logician's folly. Tech. Rep. Memo 54, Dept. Computational Logic. Edinburgh University (1972).
45. Hewitt, C. Description and theoretical analysis (using schemata) of planner: A language for proving theorems and manipulating models in a robot. Tech. Rep. AITR-258, MIT (1972).
46. Hillenbrand, T., Jaeger, A., Löchner, B. System description: Waldmeister — improvements in performance and ease of use. In Ganzinger, H. (Ed.), Proceedings of the 16th conference on Automated Deduction (pp. 232–236). no. 1632 in LNAI, Springer (1999).
47. Huang, X. Human Oriented Proof Presentation: A Reconstructive Approach. No. 112 in DISKI, Infix, Sankt Augustin, Germany (1996).
48. Hutter, D. Management of change in structured verification. In: Proceedings of Automated Software Engineering, ASE-2000, IEEE (2000).
49. Hutter, D., Stephan, W. (Eds.). Festschrift in Honor of J. Siekmann, LNAI, vol. 2605, Springer (2005).
50. Jamnik, M., Kerber, M., Pollet, M., Benzmüller, C. Automatic learning of proof methods in proof planning. The Logic Journal of the IGPL, 11(6):647–674 (2003).
51. Kirchner, H., Ringeissen, C. (Eds.). Frontiers of combining systems: Third International Workshop, FroCoS 2000 (vol. 1794). LNAI, Springer (2000).
52. Kohlhase, M. OMDOC – An Open Markup Format for Mathematical Documents [Version 1.2] (vol. 4180). LNAI, Springer (2006).
53. Lingenfelder, C. Structuring computer generated proofs. In: Proceedings of the International Joint Conference on AI (IJCAI'89) (pp. 378–383) (1989).
54. Lingenfelder, C. Transformation and structuring of computer generated proofs. Doctoral thesis, University of Kaiserslautern, Department of Computer Science (1990).
55. Lusk, E., Overbeek, R. (Eds.). Proceedings of the 9th Conference on Automated Deduction, no. 310 in LNCS, Springer, Argonne, Illinois, USA (1988).
56. Manthey, R., Bry, F. SATCHMO: A theorem prover implemented in Prolog. In E. Lusk, R. Overbreek, (Eds.), Proceedings of the 9th conference on Automated Deduction (pp. 415–434), no. 310 in LNCS, Springer, Argonne, Illinois, USA (1988).
57. McCasland, R., Bundy, A. MATHsAiD: a mathematical theorem discovery tool. In: Proceedings of SYNASC'06 (pp. 17–22). IEEE Computer Society Press (2006).
58. McCasland, R., Bundy, A., Smith, P. Ascertaining mathematical theorems. Electronic Notes in Theoretical Computer Science, 151(1):21–38, Proceedings of the 12th Symposium on the Integration of Symbolic Computation and Mechanized Reasoning (Calculemus 2005) (2006).
59. McCune, W.W. Otter 3.0 reference manual and guide. Tech. Rep. ANL-94-6, Argonne National Laboratory, Argonne, Illinois 60439, USA (1994).
60. McCune, W. Solution of the Robbins problem. Journal of Automated Reasoning, 19(3):263–276 (1997).
61. Meier, A. TRAMP: Transformation of machine-found proofs into natural deduction proofs at the assertion level. In D. McAllester, (Ed.), Proceedings of the 17th Conference on Automated Deduction, Springer, no. 1831 in LNAI (2000).
62. Meier, A. Proof planning with multiple strategies. PhD thesis, Department of Computer Science, Saarland University, Saarbrücken, Germany (2004).
63. Meier, A., Melis, E. MULTI: A multi-strategy proof planner (system description). In R. Nieuwenhuis, (Ed.), Proceedings of the 20th Conference on Automated Deduction (CADE-20) (vol. 3632, pp. 250–254). LNAI, Tallinn, Estonia: Springer (2005).
64. Meier, A., Melis, E., Pollet, M. Towards extending domain representations. Seki Report SR-02-01, Department of Computer Science, Saarland University, Saarbrücken, Germany (2002).
65. Meier, A., Pollet, M., Sorge, V. Comparing approaches to the exploration of the domain of residue classes. Journal of Symbolic Computation Special Issue on the Integration of Automated Reasoning and Computer Algebra Systems, 34:287–306 (2002).

66. Melis, E., Meier, A. Proof planning with multiple strategies. In J. Loyd, V. Dahl, U. Furbach, M. Kerber, K. Lau, C. Palamidessi, L. Pereira, Y. Sagivand, P. Stuckey, (Eds.), First International Conference on Computational Logic (CL-2000) (pp. 644–659), no. 1861 in LNAI. London, UK: Springer (2000).
67. Melis, E., Siekmann, J. Knowledge-based proof planning. Artificial Intelligence, 115(1):65–105 (1999).
68. Melis, E., Siekmann, J. Activemath: An intelligent tutoring system for mathematics. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L. Zadeh, (Eds.), Seventh International Conference 'Artificial Intelligence and Soft Computing' (ICAISC) (vol. 3070, pp. 91–101), LNAI. Zakopane, Poland: Springer-Verlag (2004).
69. Melis, E., Zimmer, J., Müller, T. Integrating constraint solving into proof planning. In H. Kirchner, C. Ringeissen, (Eds.), Frontiers of combining systems: Third International Workshop, Frocos 2000 (vol. 1794), LNAI. Springer (2000).
70. Melis, E., Meier, A., Siekmann, J. Proof planning with multiple strategies. Artificial Intelligence, 172(6–7):656–684 (2007).
71. Riazanov, A., Voronkov, A. Vampire 1.1 (system description). In R. Goré, A. Leitsch, T. Nipkow, (Eds.), Automated Reasoning — 1st International Joint Conference, IJCAR 2001, no. 2083 in LNAI, Springer (2001).
72. Schönert, M., et al. GAP – Groups, Algorithms, and Programming. Lehrstuhl D für Mathematik, Rheinisch Westfälische Technische Hochschule, Aachen, Germany (1995).
73. Siekmann, J., Hess, S., Benzmüller, C., Cheikhrouhou, L., Fiedler, A., Horacek, H., Kohlhase, M., Konrad, K., Meier, A., Melis, E., Pollet, M., Sorge, V. $\mathcal{LOUI}$: $\mathcal{L}$ovely ΩMEGA $\mathcal{U}$ser $\mathcal{I}$nterface. Formal Aspects of Computing, 11:326–342 (1999).
74. Siekmann, J., Benzmüller, C., Brezhnev, V., Cheikhrouhou, L., Fiedler, A., Franke, A., Horacek, H., Kohlhase, M., Meier, A., Melis, E., Moschner, M., Normann, I., Pollet, M., Sorge, V., Ullrich, C., Wirth, C.P., Zimmer, J. Proof development with ΩMEGA. In A. Voronkov, (Ed.), Proceedings of the 18th International conference on Automated Deduction (pp. 143–148), no. 2392 in LNAI, Springer (2002).
75. Siekmann, J., Benzmüller, C., Fiedler, A., Meier, A., Pollet, M. Proof development with OMEGA: Sqrt(2) is irrational. In M. Baaz, A. Voronkov, (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning, 9th International Conference, LPAR 2002 (pp. 367–387) no. 2514 in LNAI Springer (2002).
76. Siekmann, J., Benzmüller, C., Fiedler, A., Meier, A., Normann, I., Pollet, M. Proof development in OMEGA: The irrationality of square root of 2. In F. Kamareddine, (Ed.), Thirty Five Years of Automating Mathematics (pp. 271–314), Kluwer Applied Logic series (28), Dordrecht, Boston: Kluwer Academic Publishers, ISBN 1-4020-1656-5 (2003).
77. Siekmann, J., Autexier, S., Fiedler, A., Gabbay, D., Huang, X. (to appear) Proof presentation, In: Principia Mathematica Mechanico
78. Sorge, V. Non-Trivial Computations in Proof Planning. In H. Kirchner, C. Ringeissen, (Eds.), Frontiers of Combining Systems: Third International Workshop, Frocos 2000 (vol. 1794), LNAI, Springer (2000).
79. Sorge, V. ΩANTS — a blackboard architecture for the integration of reasoning techniques into proof planning. PhD thesis, Department of Computer Science, Saarland University, Saarbrücken, Germany (2001).
80. Sutcliffe, G., Zimmer, J., Schulz, S. Tstp data-exchange formats for automated Theorem proving tools. In W. Zhang, V. Sorge, (Eds.), Distributed Constraint Problem Solving and Reasoning in Multi-Agent Systems (pp. 201–215) Amsterdam: IOS press (2004).
81. van der Hoeven, J. GNU TeXmacs: A free, structured, wysiwyg and technical text editor. In: Actes du congrès Gutenberg, Metz, no. 39–40 in Actes du congrès Gutenberg (pp. 39–50) (2001).
82. Voronkov, A. (Ed.). Proceedings of the 18th International Conference on Automated Deduction, no. 2392 in LNAI, Springer (2002).
83. Wagner, M. Mediation between text-editors and proof assistance systems. Diploma thesis, Saarland University, Saarbrücken, Germany (2006).

84. Wagner, M., Autexier, S., Benzmüller, C. Plato: A mediator between text-editors and proof assistance systems. In Autexier, S., Benzmüller, C. (Eds.), 7th Workshop on User Interfaces for Theorem Provers (UITP'06), Elsevier, ENTCS (2006).
85. Weidenbach, C., Afshordel, B., Brahm, U., Cohrs, C., Engel, T., Keen, E., Theobalt, C., Topic, D. System description: SPASS version 1.0.0. In H. Ganzinger, (Ed.), Proceedings of the 16th conference on Automated Deduction pp. 378–382, no. 1632 in LNAI, Springer, (1999).
86. Wiedijk, F. Formal proof sketches. In S. Berardi, M. Coppo, F. Damiani, (Eds.), Types for Proofs and Programs: Third International Workshop, TYPES 2003 (pp. 378–393), LNCS 3085. Torino, Italy: Springer (2004).
87. Wirth, C.P. Descente infinie + Deduction. Logic Journal of the IGPL, 12(1):1–96 (2004).
88. Zhang, J., Zhang, H. SEM: A system for enumerating models. In C.S. Mellish, (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 298–303). San Mateo, California, USA, Montreal, Canada: Morgan Kaufmann (1995).
89. Zimmer, J. MATHSERVE – a framework for semantic reasoning services. PhD thesis, FR 6.2 Informatik, Universität des Saarlandes, Saarbrücken, Germany (2008).
90. Zimmer, J., Autexier, S. The mathserve framework for semantic reasoning web services. In U. Furbach, N. Shankar, (Eds.), Proceedings of IJCAR'06 (pp. 140–144), LNAI. Seattle, USA: Springer (2006).
91. Zimmer, J., Kohlhase, M. System description: The mathweb software bus for distributed mathematical reasoning. In A. Voronkov, (Ed.), Proceedings of the 18th International conference on Automated Deduction (pp. 138–142), no. 2392 in LNAI, Springer (2002).
92. Zimmer, J., Melis, E. Constraint solving for proof planning. Journal of Automated Reasoning, 33:51–88 (2004).