Grounded theory methodology for knowledge engineering

Brigitte Endres-Niggemeyer
Fachhochschule Hannover
University for Applied Sciences
Department of Information and Communication
Ricklinger Stadtweg 120
D-30459 Hanover, Germany
phone +49 511 92 96 606
fax +49 511 92 96 610
Brigitte.Endres-Niggemeyer@ik.fh-hannover.de

Abstract

Grounded theory methods of inductive qualitative field research have been tried and tested in the social sciences with their strong commitment to empirical techniques. They can be used in knowledge engineering for grounding systems on empirical evidence wherever the domain experts' natural way of thinking is qualitative. After an introductory explanation of grounded theory development we go through two recent studies that discover strategies – those of companies redressing their economic situation and those of expert summarizers doing their job. The work about expert summarizing is pertinent to empirical AI. Therefore I discuss it in detail. The resulting model has given rise to a small implemention. It demonstrates how expert summarization works and that grounding a summarization system on human cognitive processing should be feasible. *Grounded theories* may be of general interest to AI colleagues who want their systems to have an empirical foundation.

1 Introduction

A *grounded theory* (Glaser and Strauss 1967/80) is derived inductively from observation according to a specific research procedure. It comes into being in an environment of qualitative field research. While much empirical research starts with a set of hypotheses and tests them against empirical data, grounded theories are developed incrementally from empirical observation. A candidate concept or proposition is discovered and integrated in the emergent theory. It remains there until it is contradicted by newer and stronger evidence. If so, the new candidate is picked up. All earlier and contradicting hypotheses are reconsidered and changed or eliminated. After a laborious correction cycle through all the pre-existing theory, the researcher returns to constructing new candidate knowledge items from empirical evidence until the next incongruity comes up. The theory approaches validity as changes engendered by new data dry up. After saturation it is checked by triangulation, i.e. the systematic comparison with independent knowledge sources such as competing investigations, possibly working with contrasting presuppositions and methods.

Generally, the big effort of field research is motivated by the concern to build the grounded theory on real-world data. Second-hand information and information from ecologically disturbed settings, for instance from a laboratory instead of the natural environment, is thought to yield theories about what is studied, i.e. a second-hand view or a lab situation. This is normally not what a researcher wants.

Whereas a grounded theory can be developed without any conscious preconceptions of the researcher, this approach seems harder and more risky. It is reasonably restricted to domains that have no theory. Whoever can, applies the best domain theories of her or his knowledge to give a theoretical framework to the initial study. Thus the researcher knows better what to look for, and (s)he is better prepared to stand the long periods of uncertainty in the early phases of grounded theory building. The inductive approach implies the eventuality of contradicting the presupposed theoretical basis.

Grounded theories were conceived in the social sciences. First and foremost they apply to domains with poor theories. There, inductive theory construction is a good choice and may be the only option available. A grounded theory remains intimately related to the reality it accounts for. In knowledge engineering, we want such a theory. A cognitive or other task model must be detailed enough for implementation. So grounded theories qualify in principle for knowledge engineering tasks that can be modeled qualitatively in a natural way.

Qualitative empirical methods and grounded theories discussion is easily accessed at qualitative research websites such as Qualitative Research Links

http://www.oise.utoronto.ca/~skarsten/research/QRsites.html

and Grounded Theory Methodology On The WEB

http://www.geocities.com/ResearchTriangle/Lab/1491/gtm-19.html.

Also useful is the QualPage: Resources for Qualitative Researchers at

http://www.ualberta.ca/~jrnorris/qual.html.

The site of the Grounded Theory Institute is found at

http://www.groundedtheory.com/.

We now look at two independent investigations of human strategies made with a grounded theory approach. The first (study A) discovers 53 corporate strategies to manage a turnaround and embeds them in their environment (Pandit 1996), the second (study B) explores the cognitive strategies of expert summarizers (Endres-Niggemeyer 1998) and comes up with an intellectual toolbox of 552 strategies and a set of working steps which are managed by these strategies.

Study A is simpler in its structure. Its presents a parsimonious view of grounded theory development. After discussing some grounded theory concepts with the easier case at hand, we continue with the more complicated investigation of expert summarization. It is pertinent to empirical AI. I exemplify how qualitative field research / grounded theory development helps to ground a system empirically. This is an issue in knowledge engineering and AI in general as well as in automatic summarizing in particular.

The discussion follows the research plan of the studies.

2 Corporate turnaround strategies (study A)

Study A is a PhD thesis. Figure 1 presents its research plan.

The initial theoretical framework. The research question "What strategies do companies apply to redress their economic situation?" was drawn from the literature. With the research topic came the prior knowledge about the domain. The author uses his review of the literature as a first case study. It led to the generation of the initial theoretical framework of corporate turnaround.

The author cites Glaser and Strauss 1967/80, but the main methodological guidance comes from Strauss and Corbin 1990. Yin 1989 was used for case study methods.

The case studies. After the first case study in the literature, the truly empirical cases were selected according to the principle of theoretical sampling, i.e. because they made it possible to expand the emerging theory found in the literature.

The two real-world cases are Fisons plc, a company which experienced a turnaround in the years 1975 – 1984, and British Steel Corporation (BSC) whose turnaround happened in the period 1975 - 1989. After analyzing these two cases, the author did not learn much more by including the next case, i.e. the existing network of propositions remained stable when new data were added. He closed data acquisition. The saturation point of the theory was reached.

Field research. The field is represented in study A by the technical literature and two databases. The principal data source was archival material in the form of reports in newspapers, trade journals, government publications, broker reviews, annual company documents and press releases. The material was extracted from Reuter's Textline and Predicast's PROMT database. The author complains that it was not always complete. Sometimes, summaries replaced full text and graphics were missing. More than 100 hours of database access time were needed.

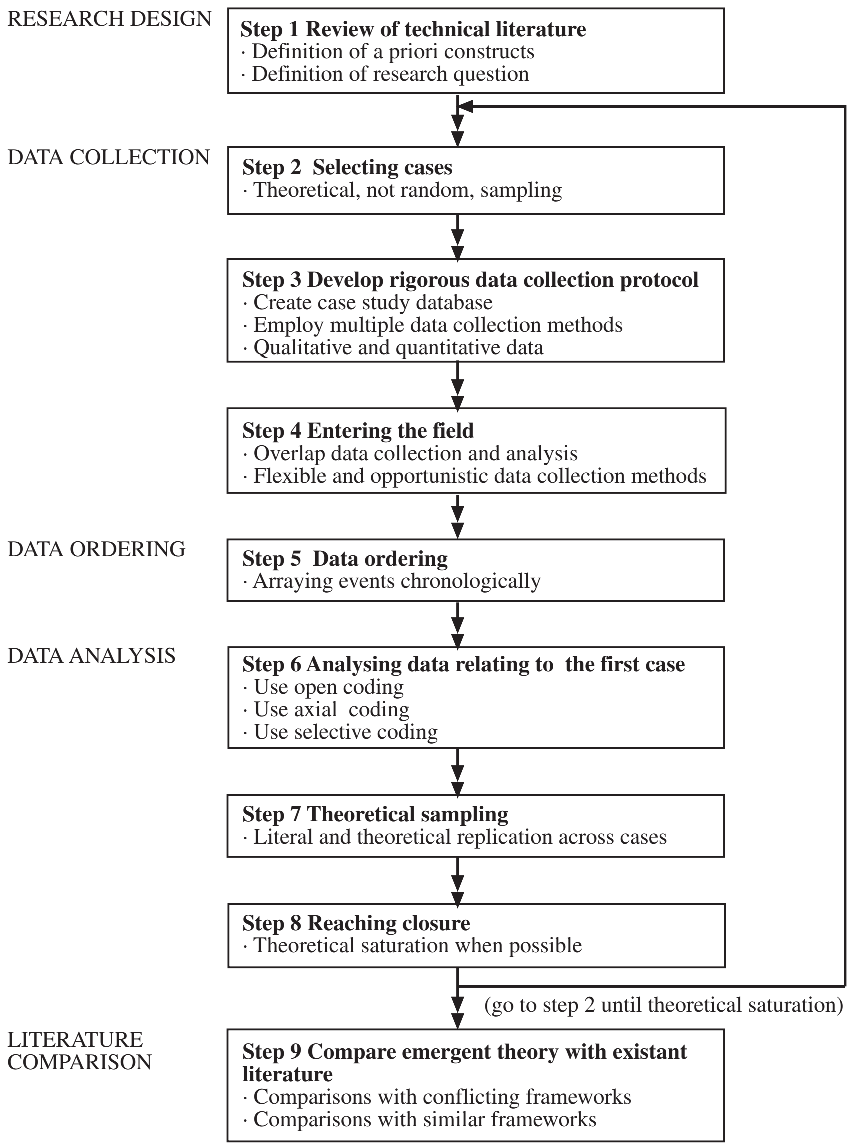

Fig. 1: Research design for corporate turnaround strategies (study A)

Information retrieval and interaction with news databases characterize the research environment in study A. The author does not report on any further contact with the studied companies. Field research remains somehow virtual and requires no particular actions. The researcher works on second-hand information. This is more practical and hard enough. It is appropriate when original information is not available or has no advantages.

Data analysis and coding. Data analysis is central in grounded theory construction. Study A follows Strauss and Corbin 1990 in distinguishing *open coding*, *axial coding* and *selective coding*.

Most important empirically is *open coding*, i.e. the labeling and categorizing of observational phenomena. Its results are concepts, the building blocks of grounded theories. Data are compared and similar incidents are grouped together under the same conceptual label.

Axial coding can be compared to classifying concepts. Main categories are connected with their subcategories.

Selective or theoretical *coding* considers observational data under categories proposed by the initial theoretical framework.

Decisions in conceptualization and coding must be recorded. This comes down to the "*memo*" event of grounded theory creation. In study A, case study databases were constructed with a software package named Atlas that specializes in qualitative data analysis. At the textual level it helped with activities such as text segmentation, coding and memo writing; at the conceptual level, it supports forming theoretical networks by activities such as interrelating codes, concepts and categories.

Evaluation. Study A includes an evaluation of the newly developed grounded theory by systematic comparison with theories described in the literature of strategic management, with the result that it resembles an earlier theory of strategic change.

3 Professional summarizing (study B)

In study B, the investigation was stimulated by teaching experience. I wondered how I could fulfil my duty of explaining professional summarizing – abstracting, indexing and classfying – to students without having a related theory, and I wondered even more that colleagues did not miss it. Figure 2 gives an overview of the research design.

3.1 The preparatory phase

The initial theoretical framework. Since Kintsch and van Dijk 1983 proposed an outline of discourse comprehension including summarization, the theory of professional summarizing I was looking for had a scientific umbrella and seemed feasible. The scientific umbrella also provided some ideas of how one might discover discourse processing strategies. They were later discarded in favor of the thinking-aloud technique for capturing data about cognitive processes. Thinking-aloud protocols are applicable for learning about summarization processes.

A detailed discussion of introspective data and methods, in particular of thinking-aloud protocols, is given in Ericsson and Simon 1980/84. Obeying the basic thinking-aloud instruction (practical formulation: "Keep talking!"), the subject talks about all operations during task execution and thus records all states of the process. According to the basic assumption of thinking-aloud data capturing, the reported processes leave their traces in the thinking-aloud record. Protocol analysis aims at explaining the observed cognitive activity by assigning it to procedures of a human cognitive program (cf. Bereiter and Scardamaglia, 1987). Subjects produce rich introspective data, but the record is also prone to gaps and distortions. Interpretation therefore draws upon additional evidence from input, output, and other sources in order to reconstruct a more complete image of the cognitive activities from their traces in the thinking-aloud protocol. Figures 3 and 5 show what a thinking-aloud protocol segment looks like.

Summarization is a rather complex intellectual process. Human cognitive processes reflect situational affordances and disturbances. If the aim is to computerize the empirical model it is normally of no interest to study disturbed expert summarizing. Therefore, the expert should stay in her or his natural environment, summarizing routine material of her or his own choice, and be comfortable during cognitive work. The more natural and original the data, the more naturalistic the grounded theory.

Diesing 1971 explains the inductive and deductive modeling which was used for the investigation of summarizing. The aim of grounded theory construction was a so-called naive model (or theory) of the process (Norman 1983). Models and theories of this type allow a subject to act purposefully in a domain, without claiming to tell the definite full truth about it. Instead they tell us enough and are close enough to the truth to enable successful action.

The KADS methodology of model-driven knowledge engineering (Schreiber et al. 1993) integrated the empirical research in an overall research process aimed at implementation. KADS divides expert system development into three phases: empirical modeling, system design, and implementation. It is up to the researcher to choose the tools to fight his or her way through the KADS procedure. Grounded theory methodology is a good fit for the empirical modeling phase. My conceptual model of professional summarizing was realized as a grounded theory.

The empirical practice followed a German leader, Mayring 1990. With him, I stuck to the full methodological requirements specified by Glaser and Strauss 1967/80. I obeyed the 13 tenets of qualitative field research:

1. *Case studies*. During the research process it is necessary to document and analyze particular cases. They are essential for checking the adequacy of procedures and interpretations.
2. *Openness*. The research process remains open to the domain of investigation, such that the theoretical structuring, individual hypotheses, and the methodology can be revised as needed.
3. *Methods control*. In spite of its openness, the research must follow a controlled procedure. The individual steps obey well-founded rules which are explained and documented.
4. *Prior knowledge*. In the domain of social sciences the analysis is always molded by the analyst's prior knowledge. Therefore her or his prior knowledge must be made explicit and developed under the influence of the investigation domain.
5. *Introspection*. Introspection is a valid source of information.
6. *Interaction between researcher and domain*. Research is not seen as the registration of so-called objective features of the domain, but as an interactive process during which both the researcher and the domain may develop and which allows for subjective interpretations to come up and to change.
7. *Holism*. Functions and contexts of human life which have been modularized for investigation must be integrated again for a holistic interpretation.
8. *Historicity*. Since all domains of the humanities have a history, the approach to them must be predominantly historical.
9. *Problem orientation*. Concrete and practical problems of the domain are preferred in research. They structure the scientific results.
10. *Generalization by argumentation*. When generalizing results of research in the humanities, an explicit argumentation must explain which results can be generalized to which situations, domains, and time periods.
11. *Induction*. In the social sciences, inductive methods for backing and generalizing results are central. They need control.
12. *Rule concept*. In the humanities, regularities are better described by context-dependent rules than by general laws.
13. *Quantification*. Qualitative investigations prepare meaningful quantifications that may support the validation and generalization of the results.

This means in research practice: Data acquisition and sampling follow a *case study* methodology. *Openness* to the field situation and *inductive reasoning* preclude imposing preconceived hypotheses, or mere hypotheses testing instead of theory generation from observation. In addition, openness of research practice pleads for methods that adapt to the object of investigation. It goes without saying that the adapted *research procedure* must be *documented* in detail because it cannot refer to a general standard in a simple way, and that it must be applied in a *controlled* fashion once it is established. For investigating cognitive processes, *introspection* is a valuable source of data. In contrast with ideas in hard science, the researcher *interacts* with the field, in particular with the field subjects or experts who contribute their knowledge. A separation of observation and interpretation is out of the question, because both researcher and field subjects have to communicate while constituting evidence. Field subjects are knowledgeable, the researcher will return to them for discussion and correction of the emerging theory. Of course, the investigator brings *prior knowledge* in, the more and the richer and more flexible, the better. Good qualitative research practice explains the theoretical presuppositions instead of negating them. As soon as a piece of theory has been generated from evidence, its scope must be established by an explicit *argumentation. Generalization* may often be limited, since we do not expect laws in social science and humanities, but rather *rules* that are subject to exceptions. From the viewpoint of research strategy, we want a *holistic* view of a phenomenon, and *real world problems* deserve priority. *Historicity* may be important, and *quantification* can follow the explorative work of grounded theory creation.

Methods planning and testing. All my data capturing happened in case studies. Phases 2 and 3 served methods planning and testing only. The small sideline study (phase 5) supported methods testing as well, but it also gave publishable results.

The first and largest case study (11 summarizing processes) was done on the researcher. Its purpose was to find out whether the strategies postulated by Kintsch and van Dijk 1983 could be pinpointed in my own brain. I started with their four macrorules as formulated by Sherrard 1989:

DELETE trivial and redundant information.
SELECT a topic sentence already in the text.
SUBSTITUTE a general term for a list of objects or a sequence of actions.
INVENT a topic sentence if it does not appear in the text.

Then grounded theory generation followed the well-known basic comparison technique. If one of these rules was applicable to what had happened in my mind, I used it. As often as I felt sure that no prefabricated macrorule applied, I defined a new one and added it to my strategy pool. Thus a starter set of strategies was generated. I put all that aside for some months. At my return, I was very astonished that my strategies had not vanished, but were still good.

I saw that 11 processes were largely enough to obtain a reasonably precise image of an individual strategy set, so the experts of my future field research were asked to perform 9 summarizing processes. From my own experience, I prepared for them instructions for producing thinking-aloud records. After that, the results were not used any more, because they had been produced at different levels of theoretical and practical research competence. Having done the thinking-aloud procedure myself helped later in convincing colleagues to participate in the field study.

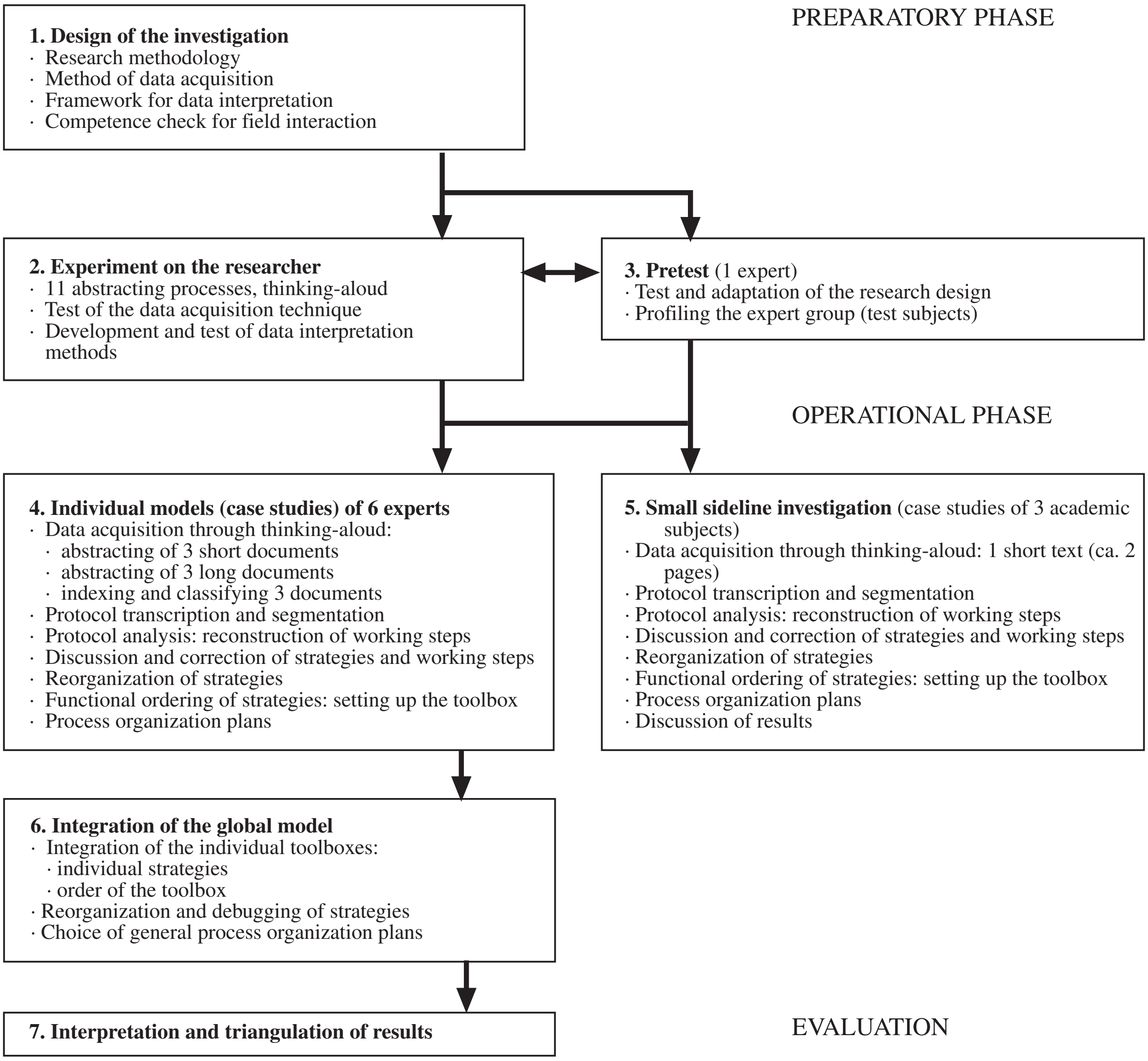

Fig. 2: Research design for professional summarization (study B)

A small preparatory study with one colleague (the pretest – phase 3 of Fig. 2) made definitively clear that thinking-aloud had to be used for data capturing, the alternative proposed by Kintsch and van Dijk 1983 was prohibitive in terms of expert time. I learned what profiles participating experts had to bring in:

- In order to control possible cultural biases they had to represent at least two different cultures, and as many scientific disciplines as possible.
- They needed some teaching experience as they would have to talk about their activities, not only to perform them. Considerable steadfastness was desired to defend their convictions against those of colleagues and possible wrong preconceptions of the investigator.
- Everybody should be able on his or her own to state the methodological knowledge of professional summarizing (abstracting and indexing) in the field.

According to case study methodology (Pappi 1987, Yin 1989), pooling the results of independent case studies strengthens the validity of a study. Case studies with half a dozen persons of the required qualification and standing were anticipated to accumulate enough competence probes for a reasonably reliable inter-individual theory. Through personal contacts I found experts who fulfilled the requirements stated above. They were certainly neither a random sample nor a theoretical one, but they could represent the methods of the field.

When beginning the big and laborious field study with the summarization experts, I tested in the small sideline study (phase 5 of the research design), whether the strategies of my mind were shared by two doctoral students. The strategies showed up with everybody, they agreed reasonably, the methodological equipment worked, and we published our results (Endres-Niggemeyer et al. 1991).

3.2 The main field study

In the fully-fledged field research of study B the researcher traveled to Los Angeles, to Rockville and College Park in Maryland, to Karlsruhe, Bonn, Trier, and Frankfurt in Germany to see the experts at their office or at home. Sometimes I participated in selecting documents for the experiment, sometimes I was present during taping sessions, sometimes taping had gone wrong, I managed circumstances and the expert's wishes. We tried together to get it alright. We discussed and corrected my interpretations, sometimes with stupendous effects, until I had learnt to meet the expert's own view.

Data acquisition by thinking-aloud. The experts taped nine summarizing processes per person with short and long documents, respectively, producing abstracts or indexes and classification notations (cf. Fig. 2, phase 4). They used 52 different documents of their own choice.

In order to prepare data interpretation, the tapes were transcribed. A simple orthographic transcription was used, recording pauses and background noises, and indicating deviations from standard English or German. Readers can inspect the style of transcription by looking at the protocol segments in Figs. 3 and 5. Next the transcribed working processes were divided into steps. Working steps are delimited periods of cognitive activity. They are very common in long cognitive efforts like summarizing because people simply need breaks. The breaks were identified in the thinking-aloud protocol by cognitive boundary signals expressed through interjections ("now let's ...", "ok", "oops", "next", "finished", etc.), pauses and ends of cognitive actions (input-output cycles, switches to new activities, dealing with new input items, etc.).

Deriving new strategies from observation. Inductive concept formation works by abstracting from data, in the present case by defining strategies from the evidence in a working step, under the constraints which are given by the theoretical framework, the whole of the observational data and the method of grounded theory development. In the case of summarizing, the data are a thinking-aloud protocol, the input document, the written output and sometimes other material, e.g. a thesaurus. The interpretation started out from the framework of Kintsch and van Dijk 1983 and operationalized its strategies to a degree that a name, a functional definition, and application environments could be given (for a more detailed explanation see below).

Fig. 3 illustrates how a protocol segment is interpreted in order to find the cognitive strategies that account for the observed activities. The researcher assigns observed cognitive activities to strategies - recurring tasks of cognitive processing - until all activities are explained. If a suitable strategy is missing, it is created. In our step Mackin-1, only frequent strategies occur, for example:

start-explore. There is the beginning of a cognitive process, expressed by "so, reading introductory material", i.e. the start signal and the following characterization of the started process. What is started is an exploration sequence made up predominantly of exploration steps. Starting exploration sequences is the task of the strategy named *start-explore.* We postulate that it is active because it is executed and reported in the thinking-aloud protocol.

explore. The current working step is dedicated to exploration. There must be a unit which directs the step to the right aim by steering the typical plan for exploration. This function is assigned to the step-leading strategy *explore.* It must be behind the scenes since we observe a successful exploration step.

plan. After starting the working process, the summarizer announces what he is going to do now (" reading introductory material"). The strategy *plan* accounts for this unit of observed behavior. Since the summarizer shows a systematic and routine approach, the *plan* strategy is assumed to simply activate and verbalize the first item of a standard working plan in memory. By explaining the strategy thus, its later implementation becomes visible.

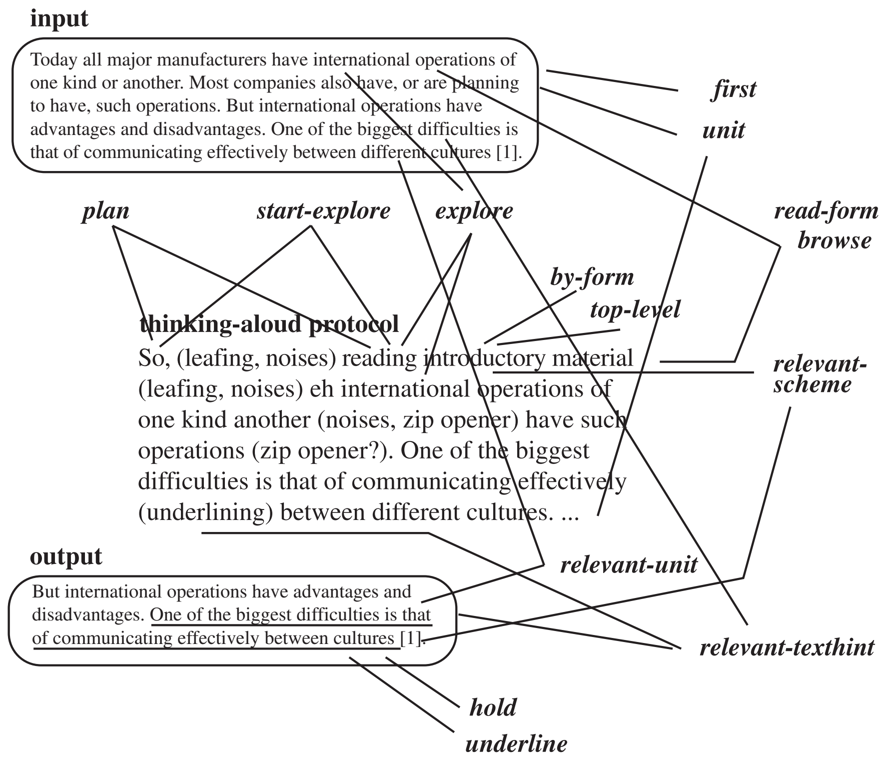

Fig. 3: Some strategies and their thinking-aloud evidence in working step Mackin-1.
Arcs connect strategies to their home text passages.

browse. The summarizer picks up information from the source paper. While thinking aloud, he reads it to the listeners. There is no evidence for any special constraints that restrict his information intentions, for instance to highlighted items. So we assume an open acquisition intention: the summarizer is reading for understanding. The strategy *browse* imposes the observed open acquisition style. Incidentally, modeling information acquisition intentions is necessary because they can vary with professional summarizers to a great extent.

read-form. There must be a visual reading process, copying characters from the source document to the brain of the summarizer. Since there are no complications of information intake in our sample step, a basic reading functionality suffices. Therefore the *read-form* strategy is assumed to work. It recognizes characters and layout features (necessary for finding the beginning of the paragraph and the text) and copies them to a memory register.

The remaining strategies of the working *step (by-form, top-level, first, unit, relevant-scheme, relevant-texthint, relevant-unit, hold* and *underl*ine) are discovered by the same technique.

The role of constraints. As grounded theories are constructed inductively, one may fear that the resulting theories are no more than the investigator's personal interpretations. This is not true, because concepts of a grounded theory are established under heavy constraints. Combined constraints largely preclude unfounded hypotheses, without excluding alternative interpretations. Besides being suggested by factual evidence, a newly constructed strategy must fulfill a set of restrictions (see Fig. 4 – the reading direction is counterclockwise). It must:

- match the assumptions of the *guiding text processing model.* If data force interpretations which are not covered by the research guiding model, and which cannot be changed, the model must be adapted. If it cannot be modified accordingly, the investigation has a serious flaw.
- be established according to the rules of *qualitative modeling methodology.*
- comply with the *local data integrity conditions*, i.e. input, output and memory contents in the current step. If for instance observed output contradicts the assumption that a specific strategy has been active, the strategy must be retracted.
- fit into the *local processing integrity conditions*, i.e. share the cognitive work with the other strategies of the step such that empirically observed behavior is achieved and the whole step can be reconstructed.
- respect the conditions set by *input, output, and memory contents* during earlier and later working steps.
- preserve the *functional unity of the strategy in all working steps* where the strategy occurs. The definition of the strategy is identical in all its contexts. If a definition is functional in one step and is not in another, the researcher must find a better solution: split a strategy in two, change its approach, make it more general or anything else that copes with all constraints.
- preserve *dissimilarity from other strategies* in the individual toolbox, avoiding an unjustified functional overlap.
- respect the *functionality of the individual model* in which it takes part. This condition would be violated for example by strategies that compete with others for the same task.
- contribute to the *functionality of the group model.* Strategies which occur twice under different names in different individual models are a frequent sin against this principle. They must be weeded out.

Some of the restrictions for strategies and theory are brought in by the case study technique. They become visible when the individual models are integrated into a group model. Then, strategies need correction if they do not comply with the functionality of the group model, e.g. if they present an uncontrolled functional overlap with another expert's strategy or an inability to cooperate in some working step of some other expert's individual model. It may also happen that individual models must be restructured in order to fit them into the classification scheme of the group toolbox.

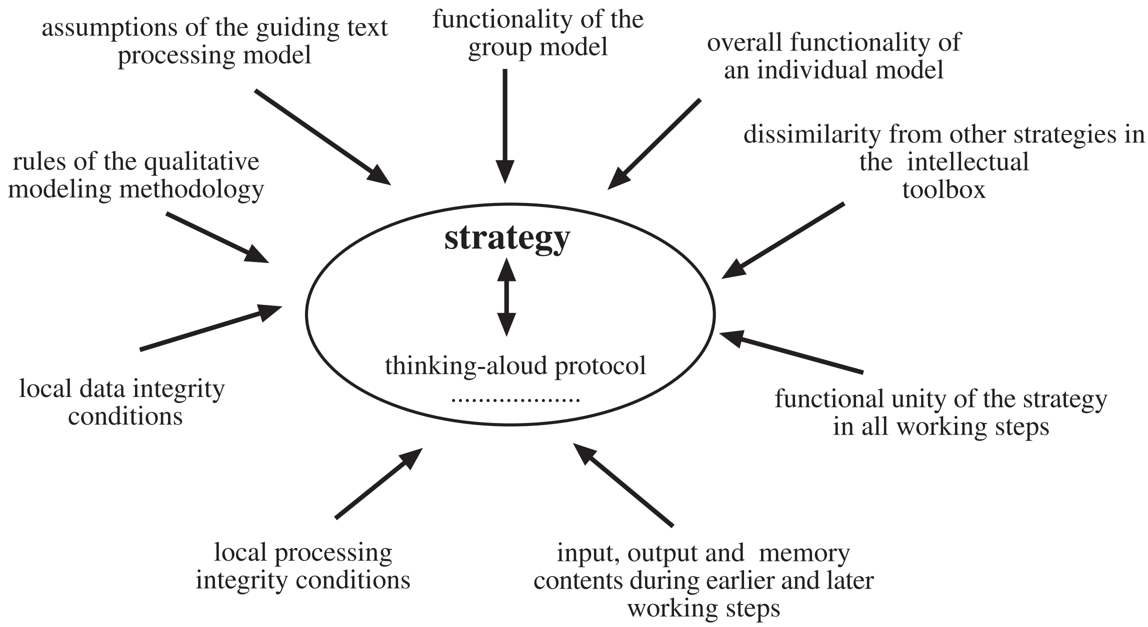

Fig. 4: Constraints on concept formation

Setting up a summarizer's individual model. While going through the summarizing processes of an expert, the occurring strategies were noted and ordered hierarchically according to their functions. The ordered collection was dubbed the expert's individual intellectual toolbox. While my documentation of the working processes was stored in traditional files, the strategies were assembled in a relational database.

To describe the structure of the working processes more globally, process overview plans were developed. Working sequences of general interest were chosen for studying the interaction of strategies.

Integrating the group model. When a group of experts solve comparable tasks, their strategies are presumed to have items in common. In the summarization study, common methods knowledge was expected to show up in the form of shared strategies. When the six individual summarization models were accumulated, this happened as anticipated. Besides this main event, the ordering of the individual toolboxes had to be modified, and a serious debugging round took place. Table 1 presents a segment of the summarizers' intellectual toolbox.

Triangulation. Results were systematically checked against all related models and systems of reading, text production, summarizing etc. I could get hold of. There were many parallels and differences, but I found no disturbing contradictions with earlier investigations.

In addition, a part of the empirical model (some 20 working steps with 79 strategies / agents) was implemented in the SimSum system (Endres-Niggemeyer 1998). A computer model presupposes a conceptual model which is detailed and precise enough to be implementable. Thus a computer simulation demonstrates a quality of the theory that complies with this demand. It checks and corrects the empirical model.

3.3 Some results

Strategies. By strategy we understand a specific cognitive process, a unit of knowledge about methods, or an intellectual tool. The intellectual tool metaphor defines professional summarizing strategies by means of their intended function. In their empirical form, strategies have a name, a short definition, and a set of natural situations where they occur. For instance, the strategy *hold* is listed in the intellectual toolbox with its name and definition as follows:

hold: Hold a meaning unit in store. (Andreas, Edward, Hanne, Harold, Inge, Marliese)

It is used by all experts of the group. It can be observed in operation in many working steps.

The empirical description of strategies prepares their realization as implemented agents. Then, the tools are ascribed activity and intentionality of their own. The strategy definition is compatible with established ideas of how to define cognitive processes (e.g. Winograd 1983) or cognitive agents: an agent must possess some program that decides about its actions, a facility for communication with other strategies, some private knowledge, and possibly an internal working memory, and it must be able to take input and to produce output, deriving its own task–oriented view of data.

4 prof A32.22: Relevance judgements based on information quality

4 prof A 32.221: Relevance judgements based on informational reliability

285	*no-doubt:* If in doubt, leave it out. (Edward, Hanne, Harold)
286	*no-publicity:* Get rid of publicity. (Hanne, Inge, Marliese)
287	*understood-only:* Leave out what you haven't understood. (Harold, Inge, Marliese)

4 prof A32.222: Relevance judgements based on information value

288	*no-truism:* Leave out what is trivial. (Andreas, Edward, Hanne, Harold, Inge, Marliese)
289	*no-void:* Leave out what lacks content. (Edward, Hanne, Harold, Inge, Marliese)
290	*own-only:* No cited results, the authors' own results only. (Hanne, Marliese)
291	*relevant-contrast:* What stands out from the rest is important. (Harold, Marliese)
292	*relevant-new:* What is new and original is important. (Edward, Hanne, Harold, Inge, Marliese)

4 prof A32.23: Relevance judgements based on factual importance

293	*last-state:* The last state of a historical development is important. (Hanne)
294	*no-reasons:* Note the fact and not the reasons behind it. (Hanne)
295	*relevant-by-fact-known:* Determine the importance of an item using your own factual knowledge. (Andreas, Edward, Harold, Inge, Marliese)
296	*relevant-by-fact:* Items that are of factual significance are important. (Andreas, Edward, Hanne, Harold)
297	*relevant-cited:* Authors who are cited are important. (Edward, Inge)
298	*relevant-causal:* In a causal (historical) development only reason and result are important. (Hanne)
299	*relevant-doc-feature:* Document characteristics are relevant. (Edward)
300	*relevant-fact:* Factual information is relevant, other information types are not. (Andreas, Edward, Hanne, Inge, Marliese)
301	*relevant-impact:* What has (practical) consequences is relevant. (Edward, Hanne)
302	*relevant-known:* What is known to be important is relevant. (Edward)
303	*relevant-result:* The result is relevant. (Andreas, Edward, Hanne, Harold, Inge, Marliese)
304	*relevant-substance:* Chemical substances are relevant. (Edward)
305	*relevant-theory:* Concepts that are important in theory are relevant. (Hanne)
306	*relevant-whole:* The whole is relevant, not its parts. (Hanne, Inge, Marliese)

Table 1: A segment from the summarizers' intellectual toolbox

The intellectual toolbox. A main component of the grounded theory of professional summarizing is the summarizers' toolbox with its 552 empirically founded strategies (see Table 1). It characterizes the methods knowledge and the personal resources implied in professional summarizing as observed during the 54 working processes of the investigation. The toolbox is arranged according to the function of strategies. For pragmatic reasons its classification is monohierarchical. It reflects standard assumptions about the functions of the strategies without precluding additional applications, just as we would normally place and expect to find a pair of scissors in a sewing box, without excluding the possibility of using them as a paperweight or hammer substitute.

Inter-individual agreement in strategy use. A substantial core of the observed strategies belong to the knowledge of the whole group: 83 strategies are used by all experts of the group, 60 strategies are shared by five experts, another 62 strategies are common knowledge of four summarizing experts, 79 strategies belong to the repertory of three of the summarization experts, 101 strategies are used by two experts, and 167 strategies are individual. Given the conditions of the experiment, this level of agreement cannot be due to the documents, because they were different for almost all working processes. There are no inexplicable differences in the number of strategies used either (between 221 and 367), and every expert has a small private stock of methods (some 10%) which serve purposes special to her or his tasks or reflect an individual working style.

	Andreas	Edward	Hanne	Harold	Inge	Marliese
Andreas	221					
Edward	170	367				
Hanne	149	209	264			
Harold	140	212	172	269		
Inge	149	206	160	168	278	
Marliese	148	187	155	159	171	255

Table 2: Number of strategies shared by pairs of experts

Table 2 lists the number of strategies used in common by two experts of the group. The general image is that everybody is linked to any colleague by a substantial number of common strategies (minimum 140, maximum 212). Since Edward contributes a high number of strategies of his own, he is able to share more of them with others than, for instance, Andreas, who himself is a thrifty strategy user. Between these two extremes, all experts have enough in common to set up a pool of shared competence.

Since the descriptive statistics of six very different experts do not differ too much, we may conjecture how a seventh expert would be integrated in the group. (S)he would contribute some 250 strategies and share some 90 percent of them with at least one other, the remainder being personal.

Reconstruction of natural summarizing steps. Grounded theory methodology admonishes the researcher to reestablish holistic ensembles with the concepts of the theory. Since the strategies have been observed in natural working steps, we can comply with this demand and reconstruct working steps with strategies. Reconstructed working steps permit us to study how strategies interact in small cognitive units.

Fig. 5 shows a sample working step. As with all working steps elaborated for the grounded theory of professional summarizing, it is characterized by an exhibit that presents the current data constellation, and a narrative of the activities during the step. To understand the display, some lengthy hints are needed:

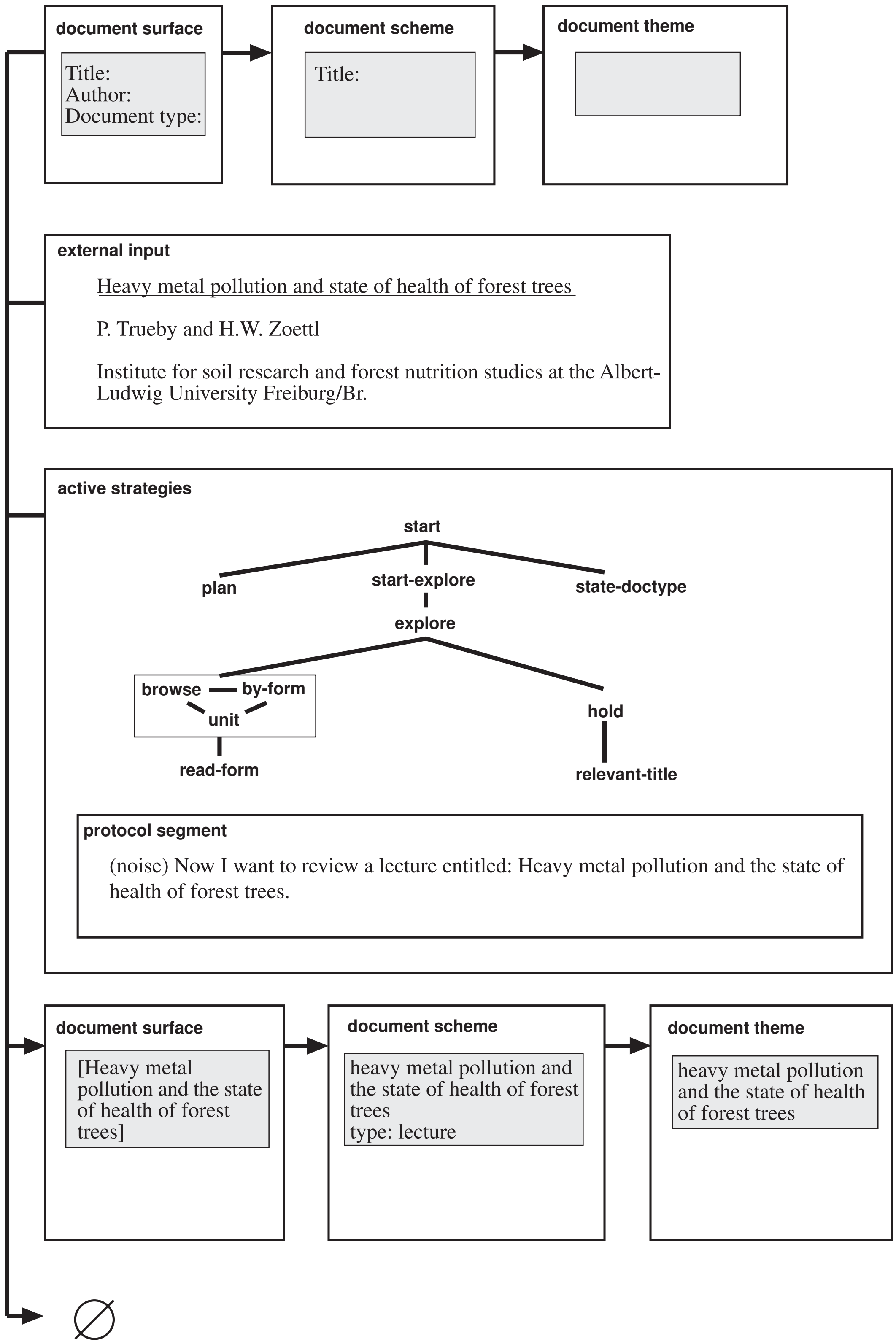

Fig. 5: Working step Trueby–1: Reading the title

Working step Trueby–1: Reading the title

Since Marliese begins a new working process in working step Trueby–1, she invests some effort in starting and initializing activity before she comes to grips with her paper. Thus we find two blocks of activity:

- *starting up*
- *exploration*

Starting up. Switching from a period of rest to a working process needs some metacognitive activity or will (strategy *start).* Marliese states her intention (strategy *plan*), namely to report a lecture. We can assume that she activates her working plan for abstracting processes. In doing so, she also determines the document type. This is done by the strategy *state-doctype*. She prepares her mind for document exploitation (strategy *start-explore*).

Processing the title. The document exploration activity is dominated by the strategy *explore.* Marliese needs no complicated reading intentions. She is guided by the document structure of a lecture (strategy *by-form*), and she accepts input in units as presented by the paper (strategy *unit)* for normal reading and understanding (strategy *browse*). The strategy *read-form* copies the title information into memory (see document surface window). Now the *hold* strategy makes the relevance assessment strategies decide whether something in the title is worth keeping for the summary. The striking argument is that the title is always relevant (strategy *relevant-title*). The *hold* strategy copies the title information into both the scheme representation and the theme representation of the source document. The working step is finished.

Fig. 5: Working step Trueby–1: Reading the title (verbal description)

In an exhibit, the state of knowledge at the beginning of the working step (the input into the process) is displayed in the top windows, the process information itself appears in the middle, and the state reached at the end of the step (its output) is presented at the bottom. The display thus shows what happens during a working step. Arrows roughly hint at the two most important reading directions. Vertically we go from the start state of the step to its end state. Horizontally, from left to right, we begin with shallow surface representations, pass through the medium-deep representation of the document scheme, and end up with the deepest representation, the document theme.

At the top of the exhibit, we normally find up to three windows representing the professional standard views on the document: the document's surface representation, the document scheme containing the outline and attached information, and the document theme where the topic structure of the document is built up during the working process.

Below the top row of schemata, the window for external input supplies the data that will be processed during the working step, i.e. the text read. There, we find underlined what the summarizer actually reads (compare the thinking-aloud protocol segment).

At the bottom of the exhibit, the task-oriented document schemata display the state of the summarizer's knowledge at the end of the working step. Sometimes written external output appears in the output window.

Summary representations show up as soon as needed. Despite being smaller in data size, the summary area mirrors the organization of the larger document representation.

Mostly, windows can only present abridged information. Gray boxes inside the windows indicate areas of activation.

It is normal to have external data mirrored by an internal representation. We assume that in the case of external input, the internal representation is the result of perception, whereas in the case of writing, the internal representation feeds the external version.

In the middle of the exhibit, the pertinent passage from the thinking-aloud protocol gives the summarizer's description of the current activities. The treelike structure above the thinking-aloud protocol gives a sketch of the cognitive apparatus (or the cognitive program system) at work behind the observational data. It explains which strategies must be active and how they cooperate to reach the aims of the current working step. All the strategies that occur are defined in the intellectual toolbox. Strategies at the root of the tree plan, restrict, and monitor, whereas at the leaves we find strategies that deal with text and its meaning. Strategies put in boxes are thought to cooperate more closely than others. The treelike structure sketches the functional structure of a working step with limited precision. Among other things, it does not explain very much in what sequence and how frequently individual strategies apply.

The structure of working steps. Figure 6 illustrates how summarizing steps are realized by cooperating strategies. A leading strategy links the current step to the working plan. It ensures that the step contributes to the overall goal, regardless of any other activities that may occur as well. When a working step serves social functions or is dedicated to planning activities, it is directly subsumed by the metacognitive self-steering component.

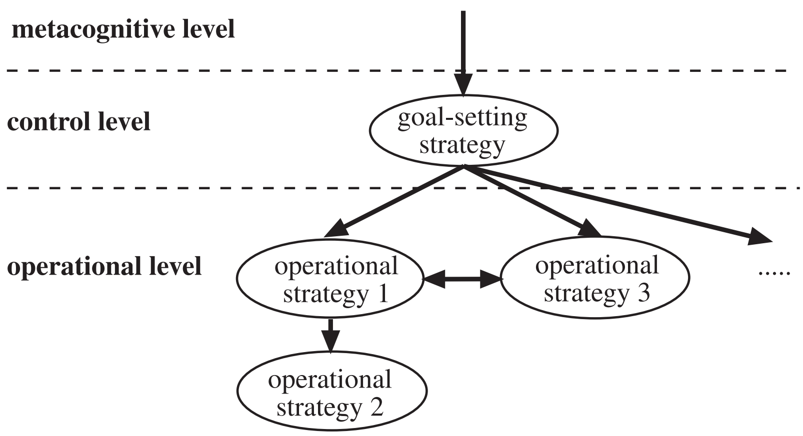

Fig. 6: Basic structure of a working step

Task-oriented memory schemata. A model of a cognitive activity must explain how the mind deals with factual data. The general explanation says that humans use schemata (Thorndyke 1979, Rumelhart 1984, Brewer and Nakamura 1984). Schemata represent objects (referents). They are reworked through thinking processes.

In the case of professional summarizing, we need three standard memory areas for document knowledge (see the exhibit of Fig. 5):

- the document surface representation
- the document scheme representation
- the document theme representation.

The three document-oriented areas combine to set up a simple mental model of the document. They are complemented by an integrated summary representation. For the sake of clarity, the three representations are separated, although the cognitively plausible assumption is that they are views which are dynamically derived from an integrated representation.

Document surface. The document surface area keeps the wording and outer appearance of the document, as far as it has been perceived. It records the surface layout of the document and its units, such as paragraphs or tables. How detailed the representation is depends on the thoroughness of perception. Representation gaps are normal.

Document scheme. Professional summarizers approach a new journal article or any other document type with some general expectations: an article will be of a certain type, it has a title, it indicates the author(s), and it somehow has an outline including an introduction and a conclusion. The expert's prior knowledge about document types and their information structure is stated by the respective document schemata. Thus the expert obtains valid expectations, which allow her or him to interpret incoming data according to the right questions, to foreground important data, background other elements, and reject anything that does not fit into the scheme. The applicable document scheme is activated at the beginning of the working process when the expert retrieves her or his professional knowledge about document representation from memory. Now the knowledge structure is prepared to accept a document outline and all knowledge directly attached to it as soon as the current document provides more detailed cues.

Document theme. The document theme area is equipped with prior knowledge about the semantic structure of objects. Conceptual relations such as *restatement* or *elaboration* provide the structural canvas of the theme representation. In practical work, we use semantic relations defined in RST (Rhetorical Structure Theory – Mann and Thompson 1988). The prestructuring of the theme representation constrains incoming knowledge to core object representations or macrostructures, starting in practice almost always with the concepts given in the document title. Semantic relations link knowledge items from the incoming document with the thematic core.

Summary representation. The summary needs a memory area of its own. In principle, it is structured like input. There must be a thematic structure (the text plan), a linear organization corresponding to the outline and a linguistic and presentational surface. In addition, an output representation must contain a working store for stages of text in progress: unformulated ideas, text under construction, planning data of a sentence, or a revised version of a text passage.

The design of real-world summarizing processes. Given meaning items and cognitive processes as basic building blocks, there are two ways to aggregate larger units: by subtasks and by meaning objects (see Figs. 7 and 8). In the first case, all items go through a specific subtask before the next subtask is begun. The summarizer submits items to a specific treatment such as reading or writing and stores them away as intermediate products, thus aggregating stages or phases of the working process. In the second case, every knowledge item is fully processed through all subtasks before the next one is picked up from the source document.

While both a clear-cut object-oriented and a clear-cut subtask-oriented process organization occur, summarizers often also blend the approaches depending on their will, the difficulty of the document or other parameters.

Phase-oriented summarizing. The more frequent organization of the summarization process goes by subtasks, having one subtask after the other deal with all available objects (see Fig. 7). First, the summarizer explores the document, storing all acquired information items in memory and possibly on an external medium by marking them and taking notes. After exploration, all results on store are assessed for their usefulness. At the end of this phase, the relevant items have been singled out. They are the next intermediate product of the summarization process, the part of the document representation that is kept for the summary. Starting from this summary representation, the abstract or indexation is produced in the next working phase. At the end, the draft may be revised. Figure 8 illustrates a working process that goes through one subtask after another. First, we have the exploration activities 1, 2, 5, 6, 7, 9, and 10. They pick up information units from the source document and deposit them in the surface representation scheme. Assessment subtasks (11–14) access these and copy relevant items to the permanent store. After that, some relevant items are noted or marked (steps 21 and 22). In the next working phase, the summary information is put together. Subtasks 31–33 assemble the material both from notes and from the memory representation. Finally, the steps 41–43 produce the written summary.

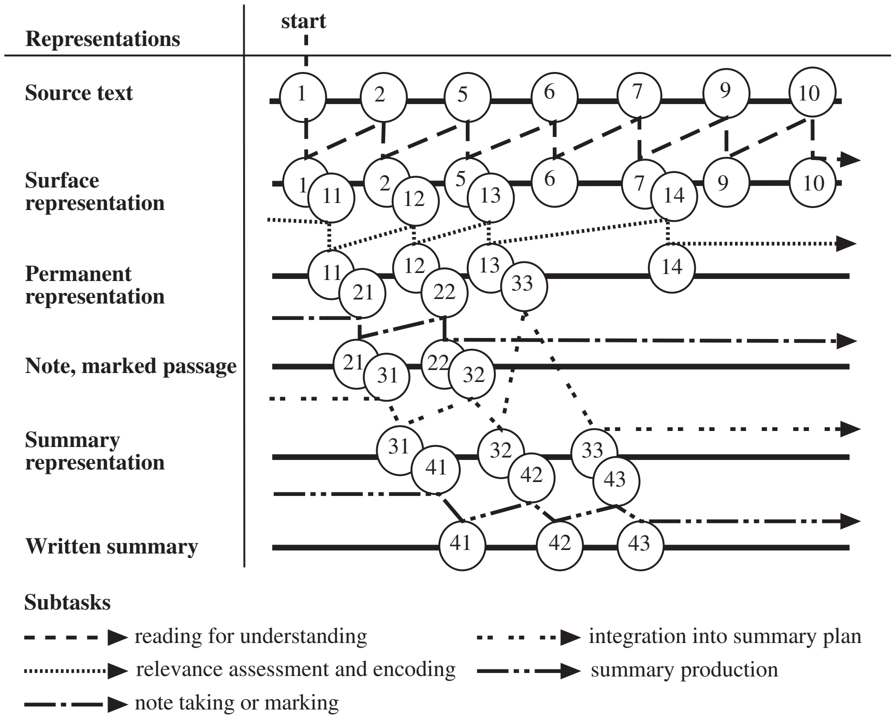

Fig. 7: Subtask- or phase-oriented organization

Production-oriented summarizing. During production-oriented summarizing, the summarizer begins on the spot to produce the target summary. Unless they are dropped, knowledge items are handed through all applicable subtasks. They are first acquired by document exploration, then assessed for their relevance, integrated in a target statement and lastly written onto an external medium such as a sheet of paper. Figure 8 shows a segment of a production-oriented summarizing process. Here, a move is responsible for an output item, integrating all necessary subtasks. It may achieve its aim or fail. In the figure, the moves 1 and 2 reach their aim using all intermediate representations. Move 3 skips the external note or marking. For lack of relevant input, moves 4, 5, and 6 do not cross the surface representation and abort, whereas move 9 is successful again.

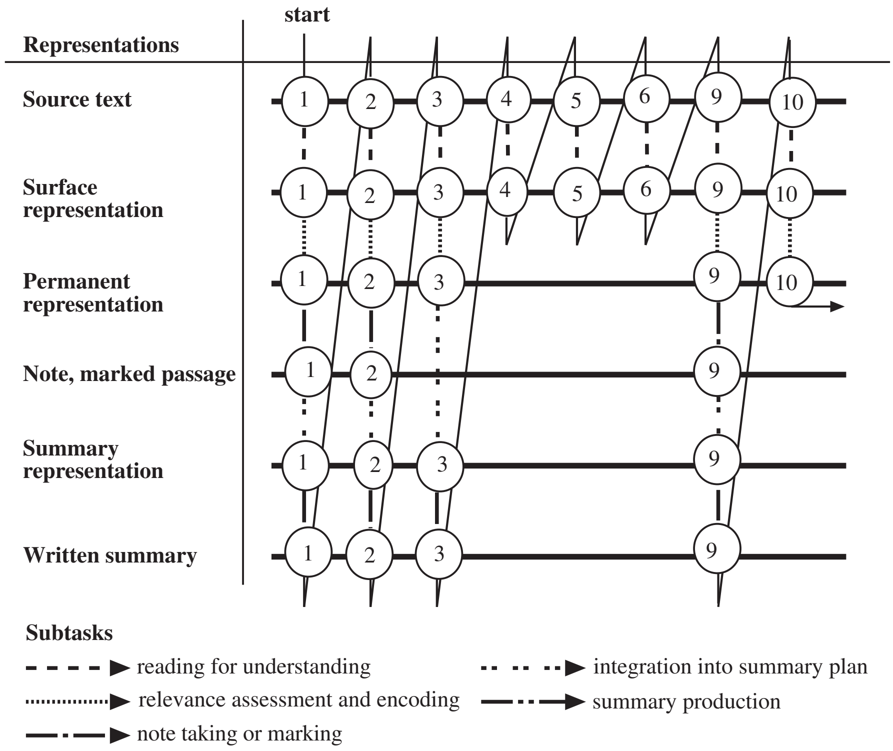

Fig. 8: Production-oriented summarization

Working plans and subplans. Since experts are experienced, they must have a repertory of ready-made working plans. The availability of these working plans ensures confident routine action in their domain. Like cooking recipes, plans state which materials and actions are needed in which sequence to obtain a result, be it a tomato soup (Hammond 1990) or a classification code. The same is true of plans of professional summarizers. They are defined at different aggregation levels, stating for instance the procedure for drawing a descriptor from the document title (an operation of moderate scale) or the strategy for abstracting a review article (a large-scale plan). Summarization experts need global overall plans and a generous number of plans for subtasks. The empirically observed subplans cover many well-known topics. Among them:

- information acquisition for an indexing term
- identifying the document topic
- recognizing an in-text summary
- writing a topic sentence
- solving an understanding problem by comparing paraphrases of the difficult passage
- obtaining abstract statements from headings
- exploiting a back-of-the-book index for indexing
- checking a synonym reference
- complementing the indexation with thesaurus concepts
- revising verb tenses
- including a superordinate concept in an indexation

4 What can been achieved with grounded theory methods in knowledge engineering and AI?

We use methods to obtain results in a controlled way. I have presented you in some detail with my grounded theory of professional summarizing. One may conclude that grounded theory methods can yield conspicuous results. The empirical results gave rise to an implementation. So we should deem them computationally useable. They convey to their implementation, the SimSum system, innovative design features: the agents representing cognitive strategies, their combined approach and blackboard communication, process organization by working steps, and so on. So an empirical grounding can make a difference.

The implementation in-the-small prepares a simulation approach to summarization for a first small real-world domain. Since an empirically founded summarizing system conforms to human argumentation in core tasks such as relevance assessment, it is assumed to produce summaries that meet human expectations better. What appears in the summary will be motivated in a human-like style, such that users can share or criticize the reasons of the system and have eventually more confidence in the information quality delivered.

Now my generalization about the value of grounded theory methods in AI: The more we want to integrate our systems into natural environments, the more they will need an empirical grounding. This argument applies to automatic summarizers, to expert systems, and to AI systems. So when preparing your next system, you may profit from considering empirical qualitative methods and perhaps grounded theories.

5 References

Bereiter, C.; Scardamaglia, M. (1987): The psychology of written composition. Hillsdale, NJ: Erlbaum.

Brewer, W.F.; Nakamura, G.V. (1984): The nature and functions of schemas. 119–160 in Wyer, R.S.; Srull, T.K., eds.: Handbook of social cognition. Hillsdale, NJ: Erlbaum,vol. 1.

Diesing, P. (1971): Patterns of discovery in the social sciences. Chicago: Aldine Atherton.

Endres-Niggemeyer, B. (1998): Summarizing information. Berlin: Springer.

Endres-Niggemeyer, B.; Waumans, W.; Yamashita, H. (1991): Modeling summary writing by introspection: A small-scale demonstrative study. Text 11(4), 523-552.

Ericsson, K.A.; Simon, H.A. (1980): Verbal reports as data. Psychological Review 87, 215-251.

Ericsson, K.A.; Simon, H.A. (1984): Protocol analysis: Verbal reports as data. Cambridge MA: MIT Press.

Glaser, B.G.; Strauss, A.L. (1980): The discovery of grounded theory: Strategies for qualitative research (11th ed.). New York: Aldine Atherton.

Hammond, K.J. (1990): Case-based planning: A framework for planning from experience. Cognitive Science 14, 385-443.

Kintsch, W. ; van Dijk, T.A. (1983): Strategies of discourse comprehension. Orlando FLA: Academic Press.

Mann, W.C.; Thompson, S.A. (1988): Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8(3), 243-281.

Mayring, P. (1990): Einfuehrung in die qualitative Sozialforschung [Introduction into qualitative social research]. Muenchen: Psychologie-Verlags-Union.

Norman, D.A. (1983): Some observations on mental models. 7–14 in Gentner, D.; Stevens, A.L. eds.: Mental models. Hillsdale, NJ: Erlbaum.
Pandit, N.R. (1996): The Creation of Theory: A Recent Application of the Grounded Theory Method. Qualitative Report 2(4), Dec. 1996. (http://www.nova.edu/ssss/QR/QR2-4/pandit.html).
Pappi, F.U. (1987): Die Fallstudie in der empirischen Sozialforschung [Case studies in empirical social research]. 365-378 in Ohe, W.v.d. ed.: Kulturanthropologie. Beitraege zum Neubeginn einer Disziplin. Berlin: Duncker & Humblot.
Rumelhart, D.E. (1984): Schema and the cognitive system. 161–189 in Wyer, R.S.; Srull, T.K., eds.: Handbook of social cognition. Hillsdale, NJ: Erlbaum, vol. 1.
Schreiber, G.; Wielinga, B.; Breuker, J. (1993): KADS. A principled approach to knowledge based system development. London: Academic Press.
Sherrard, C. (1989): Teaching students to summarize: Applying textlinguistics. System 17, 1-11.
Strauss, A. L. (1987): Qualitative analysis for social scientists. Cambridge: Cambridge University Press.
Strauss, A. L; Corbin, J. (1990): Basics of qualitative research. Newbury Park CA: Sage.
Thorndyke, P.W.; Hayes-Roth, B. (1979): The use of schemata in the acquisition and transfer of knowledge. Cognitive Psychology 11, 82–106.
Winograd, T. (1983): Language as a cognitive process. Reading, MA: Addison-Wesley.
Yin, R.K: (1989): Case study research: Design and methods. London: Sage.