Content-Based Document Retrieval Using Natural Language

Ingo Glöckner[1], Sven Hartrumpf[2], Hermann Helbig[2], and Alois Knoll[1]

[1] Technische Fakultät
Universität Bielefeld
33501 Bielefeld – Germany
{Ingo,Knoll}@TechFak.Uni-Bielefeld.de

[2] Praktische Informatik VII
FernUniversität Hagen
58084 Hagen – Germany
{Sven.Hartrumpf,Hermann.Helbig}@FernUni-Hagen.de

Abstract. A system for the content-based querying of large databases containing documents of different classes (texts, images, image sequences etc.) is introduced.[1] Queries are formulated in natural language (NL) and are evaluated for their semantic contents. For the document evaluation, a knowledge model consisting of a set of domain specific concept interpretation methods is constructed. Thus, the semantics of both the query and the documents can be interconnected, i. e. the retrieval process searches for a match on the *semantic level* (not merely on the level of keywords or global image properties) between the query and the document. Methods from fuzzy set theory are used to find the matches. Furthermore, the retrieval methods associate information from different document classes. To avoid the loss of information inherent to pre-indexing, documents need not be indexed; in principle, every search may be performed on the raw data under a given query. The system can therefore answer every query that can be expressed in the semantic model. To achieve the high data rates necessary for on-line analysis, dedicated VLSI search processors are being developed along with a parallel high-throughput media-server. In the sequel[2], we outline the system architecture and detail specific aspects of those two modules which together implement natural language search: the natural language interface *NatLink*, we performs the syntactical analysis and constructs a formal semantical interpretation of the queries, and the subsequent *fuzzy retrieval module*, which establishes an operational model for concept-based NL interpretation.

[1] Parts of the work reported here were funded by the ministry of science and research of the German state of Nordrhein-Westfalen within the collaborative research initiative "Virtual Knowledge Factory". The group developing the HPQS system includes working groups at the universities of Aachen (T. G. Noll), Bielefeld (A. Knoll), Dortmund (J. Biskup), Hagen (H. Helbig), and Paderborn (B. Monien). HPQS is an acronym for "High Performance Query Server".

[2] This is a revised version of an earlier report (Knoll et al. (1998b)).

1 Introduction

The recent advances in the fields of electronic publishing and archive technology have fostered an enormous growth in the number of documents available in electronic form. Their sheer number makes it difficult, if not impossible, to keep up with the abundance of information, i. e. to enable an effective, efficient, and accurate retrieval. Up to now, supplementing a document with textual indexing or "meta"-information, e. g. according to the "Dublin core" or "Warwick frame" standards, is the only way of characterizing documents for retrieval. Frequently, however, this meta-information is incomplete and does not lend itself to correctly identifying the stored documents the user wants to be provided with as the result of his query. Furthermore, successful retrieval based on the matching of keywords requires the user to formulate his query in a combination of words and operators, which more or less precisely circumscribe what otherwise he would have no difficulties to express in natural language. The problem of queries which use semantically adequate keywords with regard to the information wanted but not with regard to the descriptions really used with the target document, may be alleviated by expanding the single user query into many by looking for synonyms in a thesaurus. A search based on index systems or on a thesaurus results in relatively unspecific queries and hence in a very high number of matches—as painfully witnessed by users of the popular WWW search engines. Even when the user reformulates his query by excluding some of the thesaurus' entries, the number of irrelevant matches is usually high.

With non-textual documents, e. g. still images, graphics, or technical drawings, the situation is even worse: while a text normally centers around a few concepts that can be distilled into keywords, a picture normally contains a multitude of different objects. It is not the presence of the objects but very often their relations that constitute the interesting part of the image content ("A car *in front* of the house"). Moreover, depending on the context, the objects bear different names (*crowd, humans, women,* ...) that are not synonymous. In other words: For domains other than texts, it is generally very difficult to find appropriate textual descriptors (indexes) and the likelihood for retrieval misses under a given query (e. g. if objects were not registered in the index) is very high.

A small number of prototypical systems have been reported in recent literature that aim at retrieving images from a database. QBIC (see Flickner et al. (1995)) retrieves color images. It looks for regions whose characteristics can be specified by the user in terms of color, texture, and shape. It uses numerical index keys computed from color distribution and color histograms. The VIR system (see Virage Inc. (1997)) determines the similarity between images by color, composition, texture, and structure and retrieves pictures that are similar with respect to these parameters. The user can steer the retrieval process by giving individual weights to each of them. Furthermore, there have been attempts to make image contents tractable by special query languages, or to formulate the query in pictorial form, e. g. as a rough sketch. Examples of these kinds of systems are described by Bimbo et al. (1993) and Bimbo and Pala (1997); the latter gives a comprehensive list of references. Earlier work has been reported by Iyengar and Kashyap (1988).

All of these systems perform a pattern match at the level of pixels or regions. They do not aim at gaining some understanding of the image, which is necessary to handle queries relating to image content on the semantic level. In other words: they can find

small red dots in images but they do not understand the concept of a "flower". To the best of our knowledge, IRIS (see Hermes et al. (1995)), now commercially available as MediaMiner (IBM Corp. (1997)), is the only system that accepts keywords which are interpreted as concepts describing the queries for image contents. It transforms pictures into descriptive text indexes and performs the searches on these texts. Specific application domains other than searches for images of landscape type and CAD images have not been reported in the literature.

To allow queries dealing with the semantics of different document classes, to enable queries to be formulated in natural language, to exploit references between documents and fusing their information content as well as to support on-line queries obviating the need for complete pre-indexing, we have developed an architecture for a High Performance Query Server (HPQS) which is outlined in the following section.

1.1 HPQS approach

To cover the complete search cycle starting from the query input in natural language by means of recognizing the semantic content of queries and documents to postprocessing and delivery of the results, our approach combines the following methods and features into a consistent architecture:

- Formulation of queries in natural language (NL);
- Transformation of NL queries into formal database queries;
- Extraction of *semantic* information from both queries and documents (as opposed to pure keyword search);
- Fuzzy matching of query and database contents on the semantic level;
- Utilization of references between the images and the text of a single document or between documents;
- Support of run-time search on raw data (i. e. non-pre-indexed documents) using dedicated high-speed VLSI search processors.

Establishing a semantic match between the query and one or more documents in the database presupposes *meta knowledge* by means of which both documents and questions can be formally interpreted. This meta knowledge is obviously domain dependent: queries over a database of medical documents (e. g. patient files) require a model that is different from the model permitting queries over a data base of technical data sheets.

The need of a model for meta knowledge does not imply, however, that a *full understanding* of the documents is necessary. It is sufficient to know that the user query can be transformed into concepts that find their corresponding matches in the model instantiations of the documents. This obviates the need for implementing full-blown text understanding and image-understanding subsystems.

Matching concepts in NL queries and documents is not only a departure from classical deterministic and probabilistic retrieval techniques that essentially match keywords; it is the only way of unifying the different information representations within one document or across a multimedia document database. Today, such databases usually contain texts and pictures; in the near future, they will also contain audio and video signals.

Before introducing the system architecture, we take a brief look at our document domain.

1.2 Query and document domain

For the purpose of testing and practical demonstration of the system but also for guiding the design and implementation, a prototypical document domain had to be chosen carefully. After exploring the potential of several alternatives, we picked the domain of meteorological data/documents. These are available in all media of interest and can be acquired from the WWW. The range of meteorological documents comprises:

- textual weather reports (ASCII and HTML);
- time series of ground temperature and other meteorological indicators (graphics);
- data of meteorological measurement stations (tables);
- satellite images and weather maps (colour images);
- animated satellite images and TV weather reports (image sequences);
- spoken weather reports (audio).

The range of sensible queries is large, starting from simple requests for documents about the current weather, through queries about weather developments in the past fulfilling certain constraints, to questions about special weather conditions in certain locations at an arbitrary point in time. Sample queries in this application scenario include the following:

- What is the weather like in Bielefeld?
- Is it more often sunny on Crete than in southern Italy?
- Show me pictures of cloud formation over Bavaria!
- In which federal states of Germany has it been cloudy but warm last week?
- Where have temperatures been highest yesterday?
- There were how many sunny days in Berlin last month?
- Show me the average temperature readings in Germany for July!
- Show me weather maps corresponding to the current weather report!

Weather data have the advantage that they are (to some extent) comprehensible for the non-meteorologist also. The correct operation of the system may thus be verified easily even by the layman.

"Generic" approaches to content-based retrieval typically index global properties of an image (e.g. histograms). However, natural language queries like those above usually refer to specific regions of images, relationships between such regions, and accumulative properties of image regions. In the meteorology domain, for example, it is usually not interesting whether there are clouds in a satellite image, but rather where these clouds are located (perhaps even which types of clouds are located in which region). A user of the system might wish to know whether it is sunny, cloudy, overcast etc. in some city or village, which means that the descriptors in an index-based solution would need to be parametrised by geographical locations (the geographical coordinates of more than 70,000 German cities and villages are known to our system). In addition, users can be interested in accumulative properties of geographical regions (e.g. "more than r percent of Southern Germany are cloudy"), which means that an index-based system would also need pre-computed descriptors for all geographical regions of interest and all ways of expressing accumulative properties of these regions (e.g., for all choices of the above parameter r). Maintenance of such an index would not be feasible. Under these requirements, the original document seems to be the best representation of itself, provided we have a fast computational platform for online analysis.

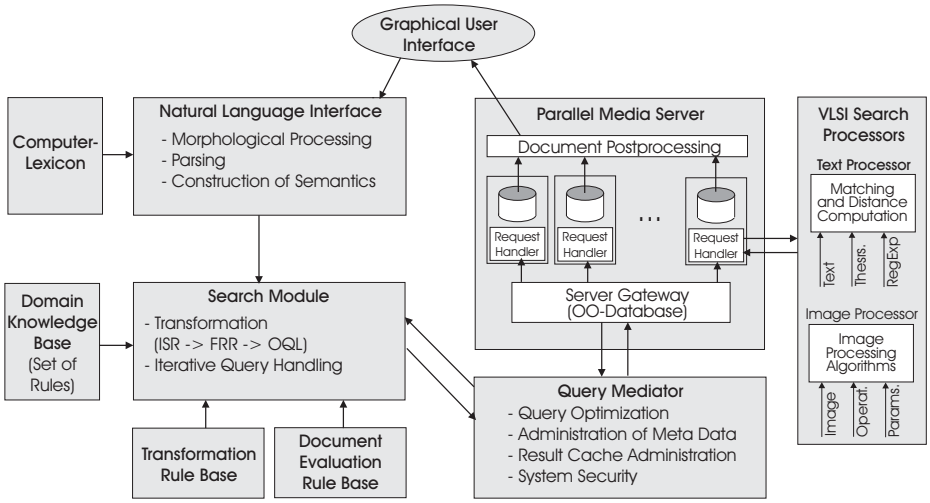
2 Overview of the system design

The user communicates with the system by means of a graphical user interface, which allows to type in queries and displays results in a reasonable, i. e. ergonomically adequate form. The query text is fed into the NL analyzer (see fig. 1), which subjects it to a syntactic-semantic transformation comprising a morphologic analysis, lexical analysis, and a syntactic-semantic parsing step for constructing a formal semantic representation of the query. The whole process is supported by a (partially domain-specific) computational lexicon. The results of the syntactic-semantic analysis are represented by means of the multilayered extended semantic network formalism (MESNET, see Helbig (1997a); Helbig and Schulz (1997); Helbig (1997b)). The MESNET graph is then transformed into the intermediate semantic representation (ISR), which contains information about query classes, speech act types, attributes, relations, local and temporal characteristics, etc.

The ISR is a purely declarative representation, i.e. not directly executable. The subsequent *search module* hence applies domain specific transformation rules which translate the declarative representation into a sequence of executable database queries. These trigger the generic evaluation and aggregation functionality of HPQS, as well as additional application methods.

To keep the interface between the database and the search module manageable, the module does not communicate with the database directly but via the Multimedia Query Mediator (MQM, Biskup et al. (1997)). The MQM accepts queries formulated in ODMG-OQL, optimizes them and invokes requests on the Media Server. Furthermore,

Fig. 1. Architecture of the HPQS



it keeps a record of meta data, handles security and user privileges, and maintains the consistency of a result cache. This caching mechanism plays a similar role in HPQS as that of pre-computed indexes in more traditional retrieval systems.

Mass data, i. e. documents and their corresponding evaluation methods, are dealt with on the Parallel Media Server. The gateway to the server essentially consists of an O_2 database that (upon requests from the mediator) generates requests to 16 parallel processor nodes, which control both the storage devices and the VLSI search processors by analyzing the requests from the gateway. Issues here are throughput optimization, load balancing, and coherency maintenance.[3] If a query requests computationally intensive operations, then a processor node forwards it to the VLSI processors, otherwise it is executed directly.

The VLSI search processors perform searches on texts. Word specifications according to the ANSI standard Z39.58 enriched by operators for approximate search are transformed into a string of processor instructions. The processor and its storage technology speed up text search by a factor of up to 1000. About the same speed improvement over software solutions is achieved by the image processors. They implement standard low-level and intermediate-level image processing techniques on color images.

HPQS offers online search and thus permits the flexible ways of retrieving and combining information necessary to support natural language queries. Apparently, a trade-off between computational effort and scalability is necessary because application of image analysis methods and elaborate text-search techniques partially contradicts with the use of a large document base. To this end, we have combined several techniques in order to ensure acceptable response times:

- *Parallelisation of method invocations*
  (always possible in our system; due to the limited number of available processing nodes, this results in a drop of processing times of about one order of magnitude);
- *Use of dedicated VLSI hardware*
  (by using dedicated VLSI hardware, the search process is accelerated by a factor of about 1000. However, this improvement is possible only for the limited number of algorithms implemented in dedicated hardware);
- *Caching of query results*
  (this technique is always applicable and yields a speed-up for frequent queries (or frequent subqueries) comparible to that of traditional indexing.

In addition, the fine-grained direct search methods supported by HPQS could be complemented with pre-computed descriptors in order to further reduce processing times, as suggested in Glöckner and Knoll (1999d); the basic idea is to use a two-stage retrieval process, in which indexing is used to prune the search space for the subsequent online-search stage.

In the following sections, we describe the design of those modules in more detail which together implement natural language access to the multimedia database, viz. the natural language interface NatLink, which generates the ISR from the NL queries, and the fuzzy retrieval module, which provides a fuzzy set-theoretic model for interpreting NL queries based on the representations generated by NatLink.

[3] A description of the job scheduling algorithm used in the HPQS system is given in Sensen (1999).

3 The natural language interface

3.1 The case for an NL interface

The acceptance of a multimedia system depends crucially on the design of its user interface. Ideally, the user interface should hide the complexity of the program, thus providing the view of an easy-to-use "information assistant". Furthermore, the interface must be well adapted to the needs of "normal users" who may be competent specialists in their field but who do not know or do not want to learn the peculiarities of formal retrieval techniques or languages. The usual way of querying information systems in terms of keywords and Boolean connectives does not meet these requirements. For the following reasons, it is not well suited to querying in a multimedia domain either:

- *Absence of keywords.* First of all, there are obviously *no words in images* being searched for, and hence keyword matching is not possible. In our framework, descriptors are assigned automatically (on demand) by means of application-dependent methods operating on the multimedia documents. These image analysis methods typically provide *much more* than keywords, e. g. descriptions of regions and local relations between regions of interest. This relational information plays an important role in image interpretation. It may make a difference to users planning their vacation in Italy whether "there are clouds" in the current weather image or whether "there are clouds in Italy". Keywords and Boolean connectives, however, are obviously not sufficient for expressing such structured, model-based information.
- *Semantic querying.* Keywords with Boolean connectives are a rather cumbersome way of specifying a user's search request. Even in the textual case, users are seldom interested in documents in which only some search keywords happen to occur. Put succinctly, texts are more to humans than a set or distribution of word occurrences; they are expressions of natural language with an associated *meaning* and informational content which may or may not fulfill the user's information needs. Ideally, users should be permitted to query an information system on the level of meaning. In other words, the whole process of querying and result presentation must be embedded in the user's system of mental concepts. In practice, it might be a long way until these ambitious goals can be accomplished. A clear advantage of using an NL interface is that it *forces* us to take seriously the issue of conceptual, semantic querying.
- *Intuitive interface.* Another problem with the usual query interfaces is their lack of user-friendliness. Although some of these interfaces (e. g. GLIMPSE, see Manber and Wu (1993)) offer powerful querying mechanisms like approximative search, regular expression search, adjacency operators, and search in specific document fields, most users are presumably not able to take full advantage of these features because they do not know when to apply them. By contrast, natural language provides an intuitive interface because everybody knows how to use his native language. Providing an NL front end not only relieves the user from learning yet another query language, but also removes technical barriers in accessing the more advanced features of an information retrieval system and makes available more information to more people in a democratic way.

Therefore, natural language queries seem to be better for information retrievel in multimedia domains than formal language queries. However, there are some possible disadvantages of such approaches:

- *Natural language ambiguity.* Automatic analysis of natural language often leads to a large set of alternatives due to (real or artificial) ambiguity in natural language or due to suboptimal analysis. If the parser chooses an incorrect alternative, the answer will probably be incorrect too and the user will be irritated or even frustrated. However, the state of the art in disambiguation advances and the natural language interface used in the HPQS contains novel, competitive disambiguation techniques (see section 3.3).
- *Computer hallucination.* Sometimes the use of natural language misleads users into asking all kinds of questions from adjacent or even unrelated domains. We hope to cope with this problem by describing the covered domain in an explicit and succinct way in the documentation and the user interface itself.

3.2 The NatLink interface

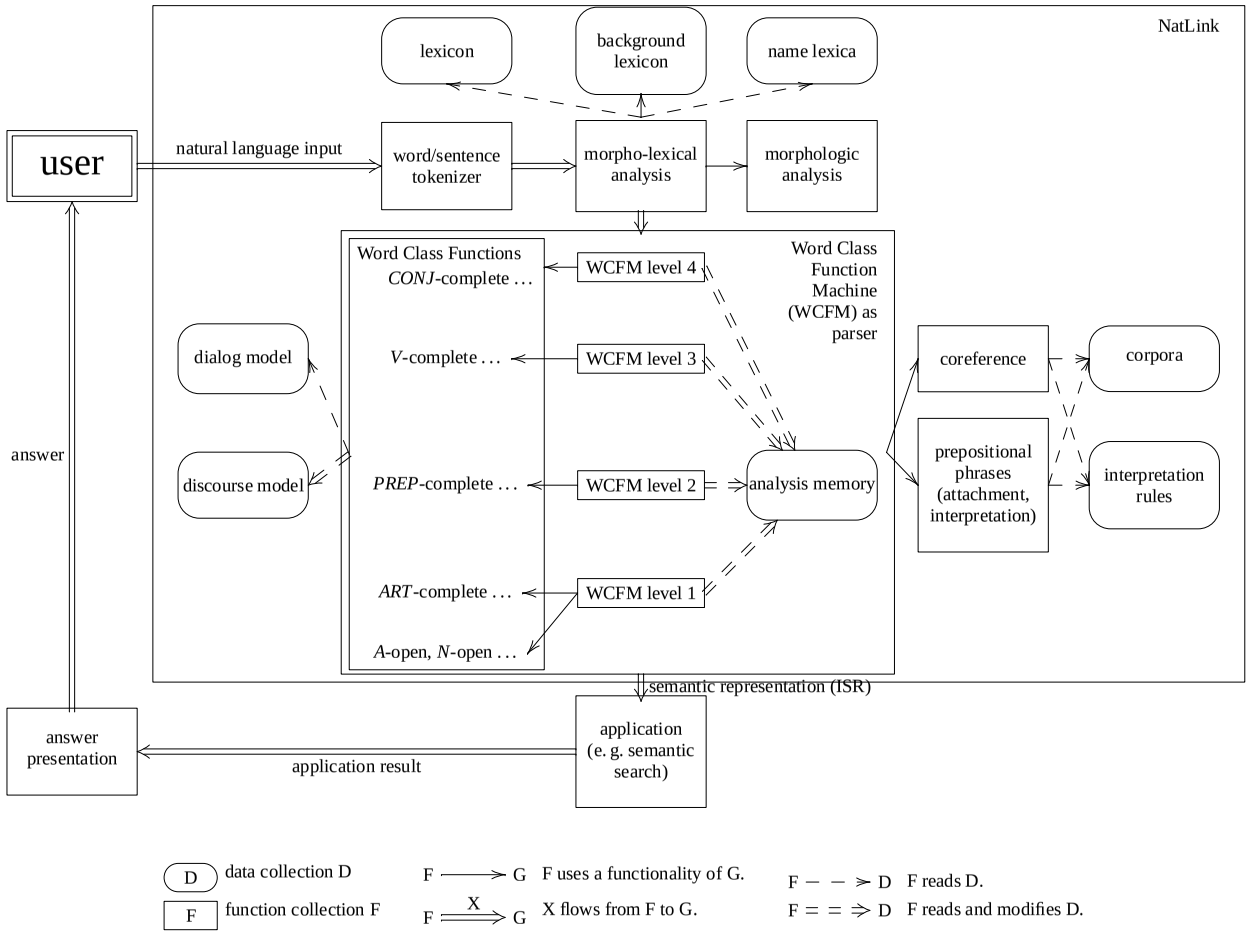
The NL interface NatLink (**nat**ural **l**anguage **in**terface to **k**nowledge) aims at the construction of adequate semantic representations for a broad class of acceptable queries as well as texts in general. In contrast to many other linguistic formalisms, emphasis is laid on the semantic acceptability of NL input, not on its grammaticality. In particular, the robustness issue plays an important role, e. g. how to cope with unknown words, elliptic sentences, and slightly ungrammatical sentences.

Robustness of NL analysis conflicts to some extent with the goal of generating expressive and deep semantic representations for NL expressions. For example, if a word is not found in the computational lexicon, morphologic analysis and syntactic context will usually provide very shallow information about the semantics of the given word. In the approach taken here, the depth of semantic analysis dynamically adapts to the syntactic and semantic information being available, resulting in a trade-off between robustness and depth of semantic representations.

A prerequisite of this approach is that the target formalism supports semantic representations on different levels of granularity or specificity. MESNET (see Helbig (1997a); Helbig and Schulz (1997); Helbig (1997b)), a multilayered extension of semantic networks, has been designed to fulfill these requirements. Due to its multidimensional structure of classificatory knowledge, it is also possible to handle generic and individual concepts as well as intensional vs. extensional aspects of meaning. MESNET has proved useful

- for semantic representation in computational lexica (Schulz (1999); Schulz and Helbig (1996)),
- as the target language for NL analysis (Helbig and Hartrumpf (1997); Helbig and Mertens (1994)), and
- as the basis for a translation into formal queries to information retrieval systems (Helbig et al. (1997, 1990)).

Fig. 2. Architecture of NatLink's WCFA implementation



The natural language input[4] is analyzed by NatLink (see fig. 2) according to the principles of the *word-class controlled functional analysis* (WCFA, see Helbig and Hartrumpf (1997); Helbig and Mertens (1994); Helbig (1986)). Like other word-oriented approaches (e. g. Eimermacher (1988), Bröker et al. (1994)), WCFA supports incremental parsing, which improves the system's robustness.

The *word class functions* (WCFs), which WCFA is based on, roughly correspond to the traditional parts of speech, but are usually more fine-grained. They comprise nouns, verbs, adjectives, adverbs, punctuation marks, determiners (e. g. interrogative determiners), pronouns (e. g. personal pronouns), etc.

All morpho-syntactic and semantic word information relevant to the analysis process is stored in the computational lexicon using typed feature structures. An example of a lexical entry for a German verb is given in fig. 3. Its subcategoriation (valence) information is semantically-oriented and includes semantic and syntactic selection restrictions for complements (e. g. *info* +, which means that the first complement of the second reading of *zeigen* must denote information) and semantic role assignments (e. g. *agt*, which means that the first complement of the first reading of *zeigen* plays the role of an *agent*). The lexicon is hierarchically structured to ensure reusability, consistency, small redundancy, and maintainability of information by means of multiple inheritance with defaults (IBL (inheritance-based lexicon formalism), see Hartrumpf (1996)). Efficient access to such lexica is described by Hartrumpf (1999b, 1997). In addition, the lexicographers can use a tool with a graphical user interface (LIA (*lexicon in action*), see Schulz (1999)), to produce new lexical entries in IBL format efficiently.

WCFA parsing is *expectation-oriented*. After morphologic and lexical processing of a word, the "opening act" of the corresponding word class function generates syntactic and semantic expectations (valency, agreement, etc.). The WCF also specifies "completing acts", which saturate or refine the expectations stipulated by the opening act. When performing a "closing act", the analysis of the current constituent is completed and the result may then be used by the WCFA parser to fill other expectations. The WCFA parser is *modularly structured* in that the process of NL analysis is decomposed into four different levels, roughly corresponding to elementary nominal phrases, complex nominal phrases, elementary propositions, and complex propositions.

Semantic representations of so-called *semantic kernels* (e. g. for noun phrases (NPs), prepositional phrases (PPs)) are constructed as soon as possible during analysis in order to resolve syntactic ambiguities in an early processing phase. Semantic kernels constitute the minimal units which a semantic representation can be assigned to in the course of incremental parsing. An example of an intermediate semantic representation (ISR) expression, which is passed on to the next component (the semantic search module) as NatLink's analysis result, is shown in fig. 4.

NatLink covers a large fragment of German syntax relevant to natural language access to data bases. The lexicon contains a considerable amount of lexemes that occur frequently in general texts plus many lexemes that occur frequently in the given application domain. By July 1999, there were 6500 lexemes with 41000 word forms in the lexicon.

[4] Currently, the natural language used is German. The examples in this paper are translated.

Fig. 3. Two simplified readings in the lexical entry for the verb “zeigen” (“show”) shown as an attribute-value matrix

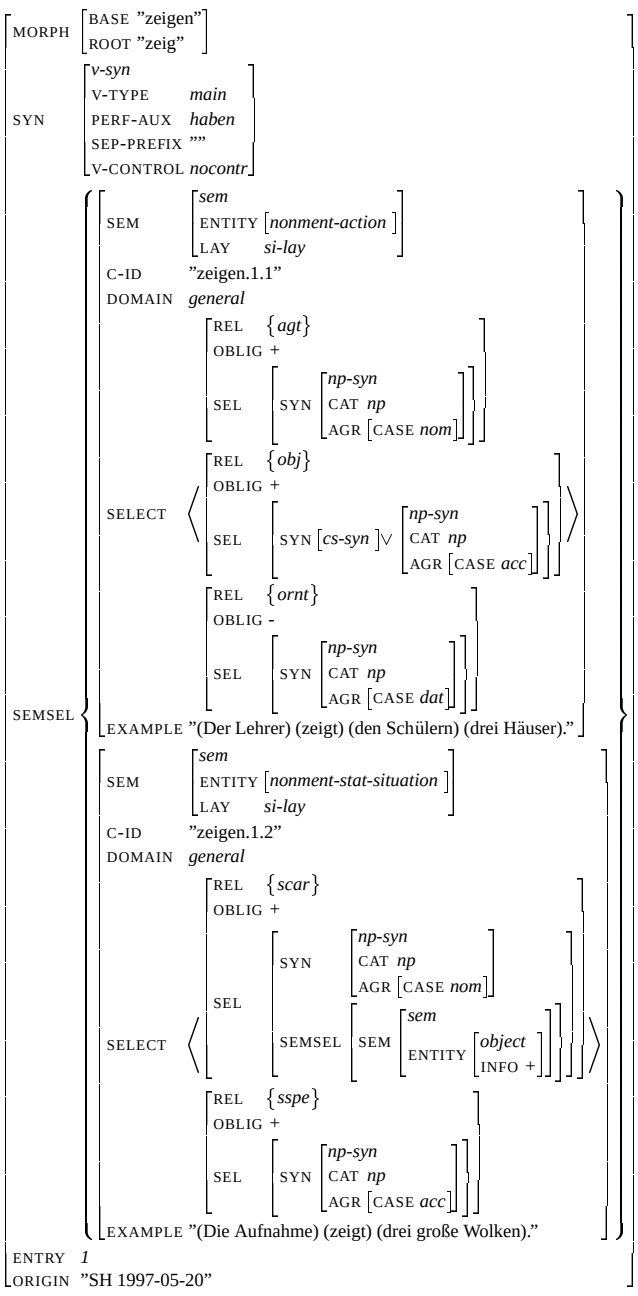
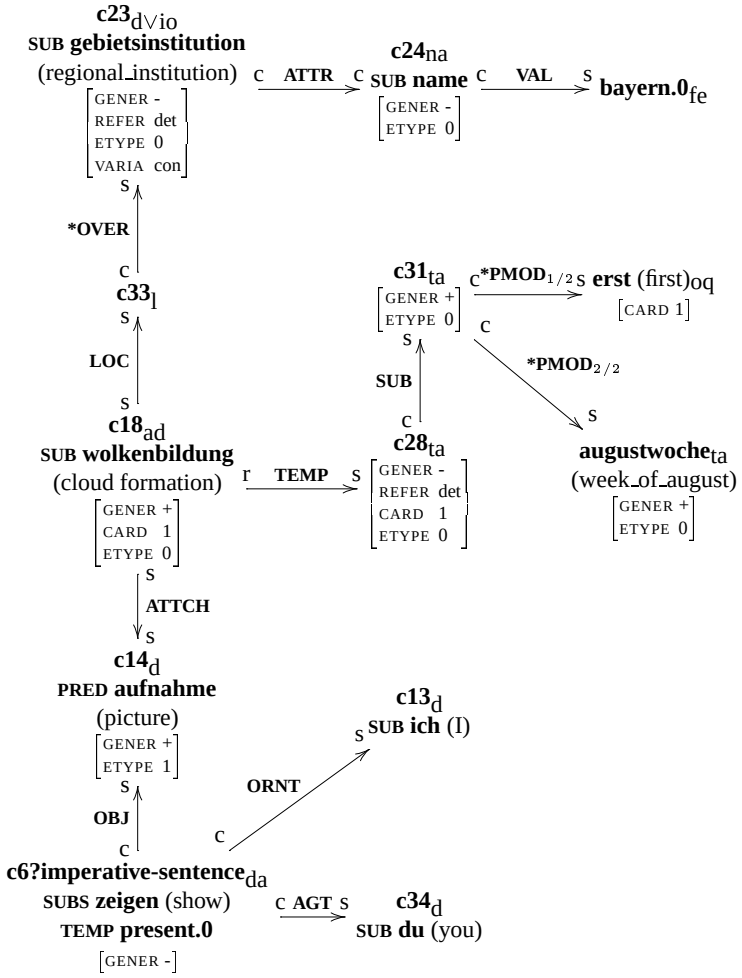


Fig. 4. Simplified ISR expression (in graphical form) generated by NatLink for the query: *Show me pictures of cloud formation over Bavaria in the first week of August!* (*Zeige mir Aufnahmen von Wolkenbildung über Bayern in der ersten Augustwoche!*)



3.3 Disambiguation in NatLink

A central problem in constructing semantic representations for NL sentences is *disambiguation*. The system should find *the* meaning the user intended and not a possibly huge set of different alternatives. NatLink uses hybrid disambiguation methods (see for example Klavans and Resnik (1996); Wermter et al. (1996)) that combines linguistic knowledge (e. g. interpretation rules for prepositions) and statistical knowledge (e. g. statistics on syntax and semantics semi-automatically generated from NL corpora). Among others, the problems of PP attachment and PP interpretation are treated this way. The disambiguation related to these two problems is described in the following (see Hartrumpf (1999a) for more details).

Ambiguity of prepositional phrase attachment is one of the most famous problems in natural language processing (NLP). In example (1) (a sentence from a German newspaper corpus), the PP *auf bayerischer Seite* can be attached to the NP *das neue Depot*, to the NP *des Deutsch-Deutschen Museums*, or to the V *gebaut* (the correct attachment place) as shown in annotation (2). The interpretation of the PP is also ambiguous. Here, the local interpretation of *auf* is the correct one.

(1) *Und wieso wird das neue Depot des Deutsch-Deutschen*
And why is the new depot the+GEN German-German
Museums auf bayerischer Seite gebaut, nachdem die Planungen für
Museum on Bavarian side built, after the plannings for
die Thüringer Talseite schon fertig waren?
the Thuringian valley-side already ready were?

'And why is the new depot of the German-German Museum built on the Bavarian side, after the planning for the Thuringian side of the valley has already been completed?'

(2) Und wieso wird das neue **Depot**$^{\mathbf{a1}}$ des Deutsch-Deutschen **Museums**$^{\mathbf{a2}}$ **auf**$^{\mathbf{auf.loc}}$ bayerischer **Seite**$^{\mathbf{s}}$ **gebaut**$^{\mathbf{m}}$, nachdem die Planungen für die Thüringer Talseite schon fertig waren?

In recent years, many statistical solutions for the problem of PP attachment have been proposed: lexical associations (see Hindle and Rooth (1993)); error-driven transformation learning (see Brill and Resnik (1994), extensions by Yeh and Vilain (1998)); backed-off estimation (see Collins and Brooks (1995), extended to the multiple PP attachment problem by Merlo et al. (1997)); loglinear model (see Franz (1996b), (Franz, 1996a, pp. 97–108)); maximum entropy model (see Ratnaparkhi (1998); Ratnaparkhi et al. (1994)).

The disambiguation method in NatLink is *rule-statistical*. It can be characterized by two key features. First, it tries to solve the PP attachment problem *and* the PP interpretation problem. Second, it is *hybrid* as it combines more traditional PP interpretation rules and statistics about attachment and interpretation in an annotated corpus. It yields results with competitive correctness for both the PP attachment problem and the PP interpretation problem.

Cross validation shows that hybrid disambiguation achieves for both problems, PP attachment and PP interpretation ambiguity, satisfying correctness results (ranges are

given for different prepositions): 88.6–94.4% for binary attachment ambiguities, 85.6–90.8% for all ambiguous attachments, and 75.0–84.2% for ambiguity degrees above 2 (leading to the multiple PP attachment problem).

Comparison of the attachment results with other approaches is possible, but difficult. One reason is that the best reported disambiguation results for binary PP attachment ambiguities (84.5%, Collins and Brooks (1995); 88.0% using a semantic dictionary, Stetina and Nagao (1997)) are for English. Because word order is freer in German than in English, the frequency and degree of attachment ambiguity is probably higher in German. There are only few evaluation results for German: Mehl et al. (1998) achieve 73.9% correctness for the preposition *mit* (*with*/*to*/...) using a statistical lexical association method.

Comparison of the interpretation results is impossible as these are the first cross-validated results for PP interpretation. But 83.3–92.5% correctness for prepositions with more than one reading seems very promising.

The high quality of disambiguation in NatLink helps to avoid frustration of users because in most cases the user's queries are understood in the intended way.

4 The Fuzzy Retrieval Module

As Fuhr (1995) observes, information retrieval (IR) differs from database querying not only in that the objects involved are far less structured, but also by the fact that with IR the relevance of an object with respect to a given query must be treated as an inherently gradual relation not expressible by Boolean logic used for database queries. With natural language input, there are additional sources of gradual information: the inherent fuzziness and vagueness of natural language concepts (cf. Pinkal (1983)) has to be dealt with on the query side ("Show me some ..." vs. unspecific queries as "What was the weather like?"; "crisp" concepts like "married" vs. fuzzy concepts like "cloudy" or "tall" which do not possess sharp boundaries). On the interpretation side, knowledge about the constituting features of the concepts in the query is often incomplete. This results in a partial mismatch between the semantic representation of the question and the target concept. Instead of postulating "strict" methods using Boolean evaluations, it seems more natural to represent such knowledge by fuzzy rules and judge the degree of applicability of these rules by means of gradual evaluations. Consequently, the well-structured view of the multimedia database system provided by the mediator component must be augmented by additional means for representing vague queries and gradual search results.

In order to to provide an adequate treatment of these pecularities of natural language queries, the retrieval component of the HPQS system utilizes a formal retrieval representation (FRR) which combines generic FRR methods (search techniques for documents of all relevant media and fuzzy methods for information combination) and domain-specific methods which implement domain concepts. The FRR functionality is implemented by appropriate classes and methods in an object-oriented database schema based on the ODMG standard, cf. Catell and Barry (1997). These methods trigger the (remote) execution of corresponding operations for media analysis and information combination on the parallel media server. The pre-defined schema can be extended by

adding application-specific classes which inherit from the generic FRR functionality. The FRR level of the HPQS system is accessed through sequences of named queries in ODMG-OQL syntax, which can invoke both the generic FRR methods and the application-specific methods. Given a user query, the retrieval component must translate the declarative ISR (as generated by the NLI) into a sequence of executable database queries which conform to the FRR schema. This translation process comprises:

- *Normalisation*:
  mapping of natural-language terms to their database correlates (e.g. names of federal states to geographic identifiers);
- *Anchoring in discourse situation*:
  (e.g. resolution of temporal deictic expressions like "today");
- *Default assumptions*:
  (e.g. in order to limit the scope of a region search to Germany).

The translation is accomplished by domain-specific transformation rules which construct the query FRR from the ISR graph. The premise part of each transformation rule specifies the structure of subgraphs to which the rule applies (e.g. temporal or local specifications, domain concepts). The consequence parts of the rules provide the corresponding FRR expressions from which the FRR of the query is constructed (Fig. 5 displays the FRR sequence generated for an example query).

Generated FRR

```
q_311: element(select x.shape from x in FederalStates
where x.name = "Bavaria")
q_312: select i from i in MeteoFranceImages
where i.date.ge(1997,8,1,0,0,0)
and i.date.lower(1997,8,8,0,0,0)
q_313: select i.pred from i in q_312 where i.pred <> i
q_314: select ImageAndRelevance(image:i,
relevance:q_311.rateGreaterEqual(0.7, i.cloudiness().
sunny().negation().germanyProjection())) from i in q_312
q_315: select ImageAndRelevance(image:i,
relevance:q_311.rateGreaterEqual(0.7,
i.cloudiness().sunny().germanyProjection()))
from i in q_313
q_316: select ImageAndRelevance(image:i.image,
relevance:i.relevance.min(j.relevance))
from i in q_314, j in q_315
where j.image = ((HpqsMeteoFranceImage)i.image).pred
q_317: select f.relevance from f in q_316
q_318: sort f in q_317 by 1
q_319: HpqsGreyValSeq(greyval_sequence:
o2_list_GreyVal(q_318)).determineThreshold()
q_320: select ImagesAndRelevance(image:f.image,
pred:((HpqsMeteoFranceImage)f.image).pred,
succ:((HpqsMeteoFranceImage)f.image).succ,
relevance:f.relevance)
from f in q_316 where f.relevance.ge(q_319) = 1
```

Fig. 5. FRR sequence generated for query: *"Show me pictures of cloud formation over Bavaria in the first week of August 1997!"*

4.1 The Formal Retrieval Representation (FRR)

The generic part of FRR pre-defines classes for the media types of interest (B/W images, RGB images, texts etc.), which provide corresponding methods for document analysis. Special emphasis has been put on fuzzy methods for information combination. The generic part of FRR offers:

- an elaborate *text-search component* (which can be efficiently supported by the dedicated VLSI processors for approximate full-text search);
- *image analysis primitives*, e.g. two-dimensional convolution, weighted median, histogram operations, morphological operations etc. (also suited for VLSI hardware support);
- *discrete and parametrized fuzzy sets* and corresponding connectives from fuzzy set theory;
- *interpolation methods*;
- *fuzzy quantifiers* which implement quantifying expressions in NL queries; these are also valuable operators for information aggregation and data fusion, i.e. for combining associated pieces of information (see below).

In addition to these primitives, we have implemented an interface to a commercial image analysis package (HORUS), which permits for the graphical development of image analysis methods.

The generic FRR can be extended by *domain-specific methods*. These have to provide an interpretation for the natural language domain concepts based on the raw document data. For example, the HPQS prototype has been taylored to the meteorology domain by implementing cartographic projections of the considered image classes (satellite images, weather maps etc.), objective ("more than 20°") and subjective ("warm") classification of temperature readings, estimation of cloud density in satellite images, classification of cloudiness degrees ("sunny", "cloudy", "overcast"...), and other domain concepts.

The implementation of domain concepts is considerably facilitated by the restriction to a specific application domain. It is the context provided by choosing an application domain which permits additional (and simplifying) background assumptions. For example, our choice of image classes (satellite images) permits the detection of clouds with relatively simple, intensity-based methods; a simplification which would not be possible with unrestricted image material (e.g., clouds on landscape photographs). In the same way that text-matching provides only a very coarse, but often still useful, approximation of text-understanding, we attempt to model only that portion of the domain concepts which must be captured to restrict the search to useful query results.

The required application methods are often rather idiosyncratic because they take into account specific domain knowledge as well as knowledge about the document types. For example, the method `cloudiness()` for computing cloudiness degrees first eliminates the contours of continents and the coordinate grid from the stored satellite images, by interpolating missing values for those pixels hidden by the contours and grid. It then uses a clustering of pixel intensities for an assignment of cloudiness degrees. Pixels corresponding to "land" and "sea" are treated differently because of the different surface characteristics. A subsequent step attempts to eliminate thin high

clouds (which do not affect the subjectively percieved degree of cloudiness) by applying a fuzzy morphological operation.[5]

Linguistic cloudiness terms like “sunny” and “overcast” are modeled by applying trapezoidal membership functions to the results of the above method. These parametric functions are provided by the generic part of FRR. Accumulative properties (“cloudy in Southern Germany”) are evaluated by applying a fuzzy quantifier, which returns a measure of the proportion of the region which is classified as cloudy. Concepts of change (“cloud formation in Bavaria”) are evaluated by combining accumulative properties of the considered region in subsequent images. With cloud formation, we have obtained satisfying results by requiring that at most 30% of the region of interest are cloudy in the first image, and atleast 70% of the region are cloudy in the subsequent image.

Let us remark that these methods can be improved (and new methods added) *at any time* because the methods are applied on-line. Unlike traditional indexing, there is no need to anticipate all potentially relevant aspects of media description when inserting the mass data.

The results of a query for *pictures of cloud formation over Bavaria* are displayed in Fig. 6; see Knoll et al. (1998a) for a detailed description of the search process.

There seems to be no generic way of supporting such idiosyncratic concepts: every system which intends to support concept-based search requires models of all considered concepts, and hence involves a considerable coding effort. In our view, the best one can do is to provide a rich set of (performance optimised) primitives for media analysis, as well as a rich set of methods for information combination. Both can be generic, but the algorithms which utilize these primitives to implement domain concepts currently must be hand-coded. In order to reduce the coding effort for domain concepts, it makes sense to utilize adaptive methods. For example, the system presented in Zhang et al. (1999) automatically learns a set of fuzzy rules which approximate an unknown target concept from a specified training sample. The resulting C-code could easily be encapsulated into FRR domain methods.

4.2 Linguistic Methods for Information Fusion

In traditional information retrieval systems, the set of operators for the aggregation of search results is essentially restricted to the Boolean connectives **and**, **or**, and **not**.[6] These connectives may, of course, also occur in NL queries, and they form an integral part of the meaning of these queries. For example, when computing search results for an NL query containing a condition “not cloudy”, the connective **not** must *not* be ignored.

The “modes of combination” of natural language, however, i. e. the various ways in which concepts might be interrelated, are by no means restricted to these connectives. In particular, the meaning of NL queries depends heavily on the *quantifying expressions* involved, as e. g. witnessed by the different meanings of “there are few clouds over

[5] for a description of methods for cloud classification, see e.g. Hutchison and Hardy (1995); Bader (1995).

[6] Adjacency (**near**) is not an aggregation operator because it applies to search terms instead of logical evaluations.

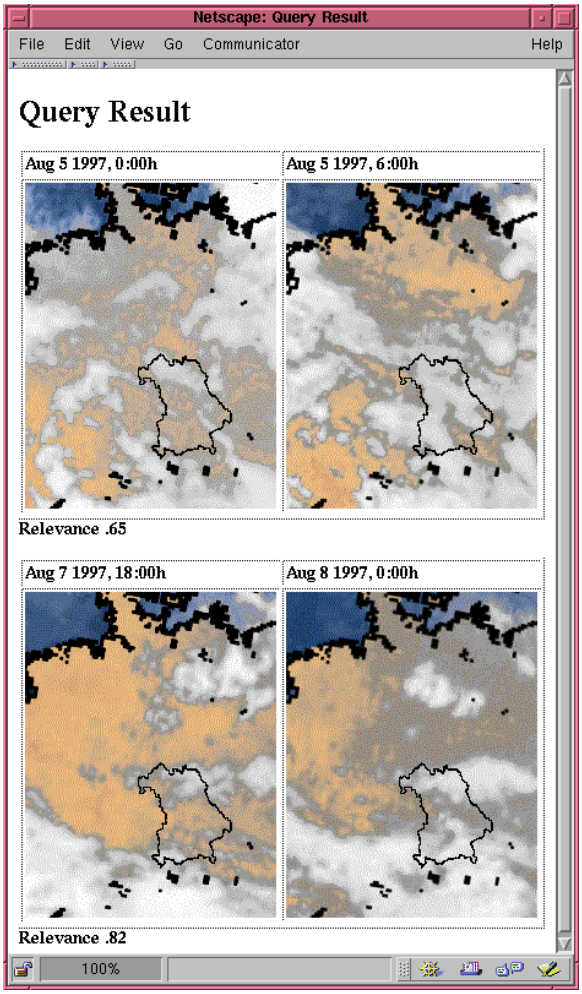


Fig. 6. Pictures of Cloud Formation over Bavaria in the First Week of August 1997

Italy" vs. "there are a lot of clouds over Italy", which both could be part of queries addressed to our system.

Because of its wealth of information combining operators, natural language access to a multimedia retrieval system cannot be provided by merely adding an NL frontend to an existing retrieval "core": in addition to the Boolean connectives, the retrieval model must also capture the meaning of quantifying expressions if content-based retrieval, which reflects the semantics of NL queries, is to be achieved (cf. Glöckner and Knoll (1997)). Examples of explicitely or implicitely quantifying expressions of relevance to

the meteorological domain are shown in Table 1.
Let us note that many of these expressions are *approximate* or *fuzzy* in nature:

Quantification over local regions
few clouds over Italy
many clouds over southern Germany
more clouds over Spain than over Greece
cloudy in Northrine-Westphalia (implicit)
Quantification over regions in time
almost always cold in the last weeks
more often sunny in Portugal than in Greece
hot in Berlin in the previous week (implicit)

Table 1. Examples of quantifying queries in the meteorology domain

- **often**, **rarely**, **recently**, **mostly**, **almost always** , . . . (temporal)
- **almost everywhere**, **hardly anywhere**, **partly** , . . . (local)
- **many**, **few**, **a few**, **almost all**, **about ten**, **about 40 percent**, . . . (approximate specification of the cardinality of a set, or a proportion of cardinalities).

These expressions are best modeled as resulting in gradual evaluations. Other natural language quantifiers have a precise, "crisp" meaning, but still we have to employ fuzzy set theory in order to allow for their application to fuzzy arguments. For example, **everywhere**, **nowhere**, **always**, **ten times**, **at least ten**, **all**, **less than 20**, . . . can be adequately modeled in the framework of classical (two-valued) logic. The query, however,

Is the weather fine in all of upper Bavaria?

requires the two-valued quantifier **all** to be applied to the fuzzy regions **upper Bavaria**, **fine weather**. In order to do so, the semantics of **all** must be generalised to the case of fuzzy argument sets.

In order to handle such cases, Zadeh (1979, 1983) has initiated research which tries to model natural language quantifiers by operators called "fuzzy linguistic quantifiers". Several classes of operators have been proposed as properly representing the phenomenon of "approximate" or fuzzy NL quantification (a survey is provided in Liu and Kerre (1998)), but there is no consensus about the proper choice, and notes on implausible behavior of these approaches are scattered over the literature, e.g. Dubois and Prade (1985); Ralescu (1986); Yager (1993) and Ralescu (1995).

The basic problem associated with these approaches is that of *linguistic adequacy*: are these approaches able to model all natural language quantifiers of interest, and do they provide *good* models which capture the intended meaning of these quantifiers? This issue of linguistic adequacy is of particular importance to a system which supports natural language querying because acceptable retrieval results will only be obtained if the system interprets the involved quantifiers (say **most**) in the way expected by the user.

A suitable framework for evaluating the linguistic adequacy of these approaches is provided by the Theory of Generalized Quantifiers (TGQ), which has developed important concepts for describing and interrelating the meanings of natural language quantifiers, see e.g. Barwise and Cooper (1981); Benthem (1983, 1984) and Keenan and Stavi (1986). Based on TGQ's distinction of classes of natural language quantifiers, Glöckner (1997) has shown that existing approaches to fuzzy quantification are too limited to cover the quantificational phenomena of interest. In Glöckner (1999); Glöckner and Knoll (1999b,a), we have conducted a formal evaluation of existing approaches to fuzzy quantification based on their compatibility with the linguistic facts as known from TGQ. The findings of this investigation indicate that existing quantifiers fail to provide good interpretations of natural language quantifiers even in the limited range of phenomena they intend to model.

We will now give some illustrative examples which reveal the failure of these approaches to provide adequate models of NL quantifiers. The examples are taken from the following *image ranking task*.

The HPQS system must be capable of ranking satellite images according to accumulative criteria such as *"almost all of Southern Germany is cloudy"*. In this image ranking task, we have a nonempty set E of pixel coordinates. Each pixel $e \in E$ has an associated relevance $\mu_{X_1}(e) \in \mathbf{I} = [0,1]$ with respect to the ranking task, which in this case expresses the degree to which pixel $e \in E$ belongs to Southern Germany, and each pixel has an associated evaluation $\mu_{X_2}(e) \in \mathbf{I}$ which expresses the degree to which the pixel is classified as cloudy (see Fig. 7). The mappings $\mu_{X_1}, \mu_{X_2} : E \longrightarrow \mathbf{I}$

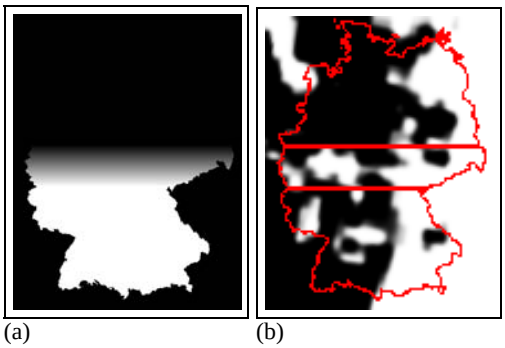
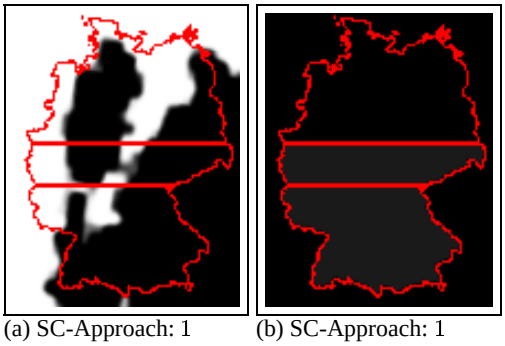


Fig. 7. Images for ranking task. (a) A possible definition of $X_1 =$ southern_germany; pixels with $\mu_{X_1}(e) = 1$ depicted white. (b) Fuzzy image region $X_2 =$ cloudy; pixels classified as cloudy depicted white. The contours of Germany, split in southern, intermediate and northern part, have been added to facilitate interpretation.

can be viewed as membership functions representing fuzzy subsets $X_1, X_2 \in \widetilde{\mathcal{P}}(E)$ of E, where $\widetilde{\mathcal{P}}(E)$ is the fuzzy powerset of E. Our goal is to determine a mapping $\widetilde{Q} : \widetilde{\mathcal{P}}(E) \times \widetilde{\mathcal{P}}(E) \longrightarrow \mathbf{I}$ which, for each considered satellite image, combines these data (fuzzy image regions X_1, X_2) to a numerical result $\widetilde{Q}(X_1, X_2) \in \mathbf{I}$ as requested

by the NL expression “almost all”. Of course, we are interested in other quantifiers also. The images are then presented in decreasing order of relevance with respect to the search criterion.

Let us now consider some results obtained from existing approaches to fuzzy quantification. Figs. 8 and 9 reveal some pecularities of the Σ-count approach, introduced by Zadeh (1979, 1983).

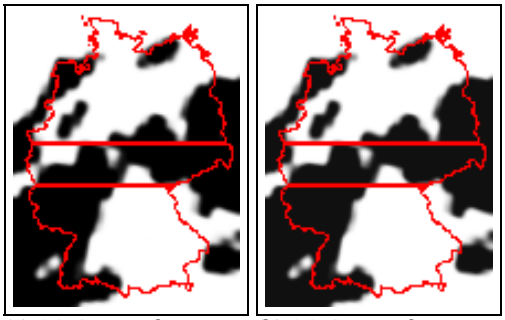


(a) SC-Approach: 1 (b) SC-Approach: 1

Fig. 8. About 10 percent of Southern Germany are cloudy (Sigma-Count)

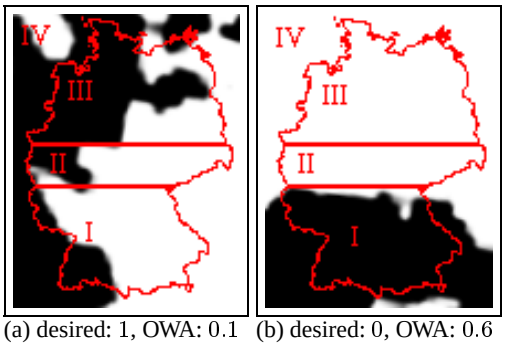
In Fig. 8, the Σ-count approach is used to evaluate the condition that “about 10 percent of Southern Germany are cloudy”. Both cloudiness situations (a) and (b) are considered cloudy to a degree of one (fully true) if the Σ-count approach is used to evaluate “about 10 percent of Southern Germany are cloudy”. While a result of 1 is plausible in case (a), this is not the case in situation (b), in which *all* of Southern Germany is cloudy to a very low degree (viz. one tenth), which certainly does not mean that one tenth of Southern Germany is cloudy. Another weakness of the Σ-count approach becomes apparent when we consider “crisp” natural language quantifiers like **all** or **atleast r percent**. In this case, the Σ-count approach generally produces two-valued results and hence cannot determine a useful ranking of the documents of interest (there are only two resulting classes of documents: fully relevant and fully irrelevant, which is too coarse a distinction of relevance for our purposes). In addition, the operators obtained from the Σ-count approach are *discontinuous* in this case, i.e. very slight changes in the membership degrees of the arguments can drastically change the quantification result. An example which demonstrates this sensitivity to noise is shown in Fig. 9. In this case, image (b) is a slightly modified version of image (a) – all pixels with a cloudiness degree of 0 have been set to a slightly higher degree. Although the difference is small and hardly visible, the Σ-count approach “jumps” from 0 to 1 when we move from the cloudiness situation (a) to the slightly modified situation depicted in (b).

Yager (1988, 1991) proposes an approach to fuzzy quantification based on Ordered Weighted Averaging (OWA) operators. Fig. 10 shows some results of using the OWA approach for evaluating the condition that “at least 60 percent of Southern Germany are cloudy”. In situation (a), we expect the result 1, because sufficiently many pixels which



(a) SC-Approach: 0 (b) SC-Approach: 1

Fig. 9. At least 60 percent of Southern Germany are cloudy (Sigma-Count)



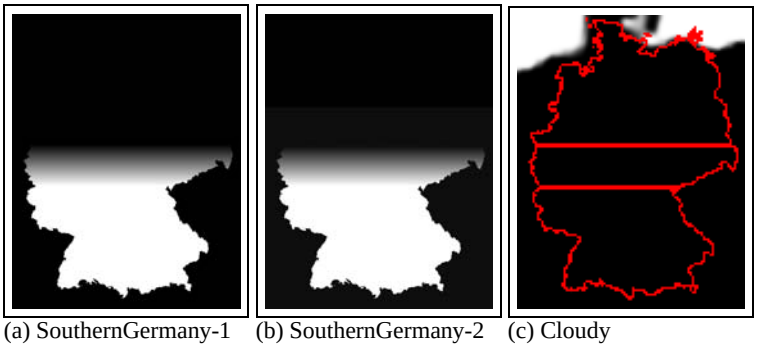
(a) desired: 1, OWA: 0.1 (b) desired: 0, OWA: 0.6

Fig. 10. At least 60 percent of Southern Germany are cloudy (OWA)

fully belong to Southern Germany (I) are classified as fully cloudy that, regardless of whether we view the intermediate cases (II) as belonging to Southern Germany or not, its cloud coverage is always larger than 60 percent. Likewise in (b), we expect a result of 0 because regardless of whether the pixels in (II) are viewed as belonging to Southern Germany, its cloud coverage is always smaller than 60 percent. OWA, however, ranks image (b) much higher than image (a). This counterintuitive result is explained by OWA's lack of a property known as *conservativity*: the cloudiness degrees of pixels in areas (III) and (IV), which do not belong to Southern Germany at all, still have a strong (and undesirable) impact on the computed results.

In Zadeh (1983) an alternative approach to fuzzy quantification is introduced, based on the concept of FG-count, which models the cardinality of a fuzzy set as a fuzzy subset of the non-negative integers. (Yager, 1991, p.72) proposes a weighting formula which generalizes the FG-count approach to the case of two-place proportional quantification which we need in the HPQS application. Fig. 11 displays some results obtained

from the FG-count approach. Let us first consider the situation where X_1 is our standard



(a) SouthernGermany-1 (b) SouthernGermany-2 (c) Cloudy

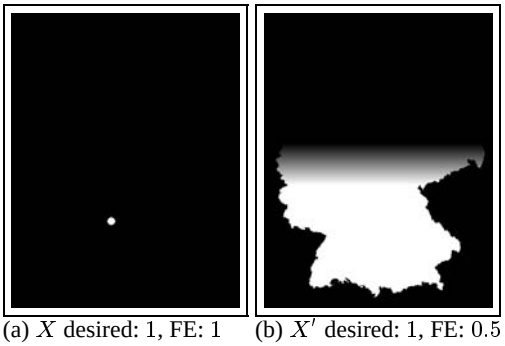
Result for X_1 = SouthernGermany-1:	0.55
Result for X_1 = SouthernGermany-2:	0.95
Desired result:	0

Fig. 11. At least 5 percent of X_1 are cloudy (FG-Count)

choice for interpreting Southern Germany (depicted in 11(a)) and where the cloudiness degrees are given by (c). We expect a result of zero because there are no clouds in the support of X_1; nevertheless, the FG-count approach produces the result 0.55. This counterintuitive result is caused by the failure of Yager's weighting formula to preserve local monotonicity properties of quantifiers.

In addition, the operators obtained from the FG-count approach can be discontinuous in the membership degrees of the argument sets. In our above example, the sensitivity to slight changes in X_1 becomes apparent when the standard definition of the fuzzy region "Southern Germany" (a) is replaced with the slightly modified SouthernGermany-2 depicted in (b). The image has been obtained from (a) by replacing black pixels ($\mu_X(e) = 0$) in the lower two thirds of (a) with a slightly lighter black, $\mu_X(e) = 0.05$. Because the upper third of the image is unchanged, i.e. the clouds in the northern part of (c) are still outside the support of SouthernGermany-2, we expect a result of 0 in this case, too. The FG-count approach, however, produces the result 0.95. The large change in the result, although we have modified our representation of Southern Germany only very slightly, shows that the operators obtained from the FG-count approach can be indeed very brittle.

Ralescu (1986) proposes to interpret fuzzy quantifiers based on the FE-count, an alternative measure of fuzzy cardinality introduced by Zadeh (1983). A definition has been given only for one-place absolute quantifiers, i.e. the FE-count approach does not provide an interpretation of the two-place proportional quantifiers needed in the HPQS system. The example depicted in Fig. 12 illustrates that the FE-count approach fails to provide an adequate interpretation of natural language even in the very restricted range of quantifiers it intends to model. The condition to be evaluated is that of determining



(a) X desired: 1, FE: 1 (b) X' desired: 1, FE: 0.5

Fig. 12. X is not empty (FE-Count)

the degree to which a given fuzzy image region is non-empty.[7] The result in case (a) is satisfactory because X is a crisp non-empty subset of the set of pixel coordinates E. The fuzzy image region X' depicted in (b) is much larger and contains X; we hence expect that if X is nonempty (which it is), then X' is nonempty, too. The FE-count approach, however, evaluates X' as nonempty only to the degree 0.5, which is clearly implausible.

These examples demonstrate that existing methods of fuzzy set theory fail to produce acceptable results in the image ranking task. A new approach to fuzzy quantification had to be developed in order to allow for an adequate processing of quantifying queries in the HPQS system. Building on TGQ, Glöckner (1997) has formulated a set of axioms which characterises "reasonable" approaches to fuzzy quantification. Every model of these axioms, called a DFS (determiner fuzzification scheme), is guaranteed to avoid counter-intuitive results like those presented above from the outset. The DFS actually used in the HPQS prototype is

$$\begin{aligned}
\mathcal{M}(Q)(X_1,\ldots,X_n) &= \int_0^1 Q_\gamma(X_1,\ \ldots,\ X_n)\,d\gamma \\
Q_\gamma(X_1,\ \ldots,\ X_n) &= \mathrm{m}_{\frac{1}{2}}\{Q(Y_1,\ \ldots,\ Y_n) : Y_1 \in \mathcal{Y}_1^\gamma,\ \ldots,\ Y_n \in \mathcal{Y}_n^\gamma\} \\
\mathcal{Y}_i^\gamma &= \{Y \subseteq E : (X_i)_\gamma^{\min} \subseteq Y \subseteq (X_i)_\gamma^{\max}\} \\
(X_i)_\gamma^{\min} &= \begin{cases} (X_i)_{\geq \frac{1}{2}+\frac{1}{2}\gamma} & : \ \gamma \in (0,\,1] \\ (X_i)_{>\frac{1}{2}} & : \ \gamma = 0 \end{cases} \\
(X_i)_\gamma^{\max} &= \begin{cases} (X_i)_{>\frac{1}{2}-\frac{1}{2}\gamma} & : \ \gamma \in (0,\,1] \\ (X_i)_{\geq \frac{1}{2}} & : \ \gamma = 0 \end{cases}
\end{aligned}$$

[7] this corresponds to existential quantification, i.e. to the condition that there exists at least one pixel which belongs to the specified region.

where $Q : \mathcal{P}(E)^n \longrightarrow \mathbf{I}$ is a compact description of the desired operator as a so-called "semi-fuzzy quantifier", $\mathcal{M}(Q) : \widetilde{\mathcal{P}}(E)^n \longrightarrow \mathbf{I}$ is the corresponding fuzzy quantifier, $X_1, \ldots, X_n \in \widetilde{\mathcal{P}}(E)$ are fuzzy subsets of E, $\mathrm{m}_{\frac{1}{2}}$ is fuzzy median[8], $(X)_{\geq\alpha} = \{x : \mu_X(x) \geq \alpha\}$ is α-cut and $(X)_{>\alpha} = \{x : \mu_X(x) > \alpha\}$ is strict α-cut. It can be shown that the integral is well-defined, regardless of Q and $X_1, \ldots, X_n \in \widetilde{\mathcal{P}}(E)$. The resulting operators can be implemented efficiently; an algorithm for the histogram-based evaluation of the fuzzy quantifiers of interest is presented in Glöckner et al. (1998).

In the HPQS system, we are currently using these operators to aggregate over fuzzy sets of pixels (local quantification) or fuzzy sets of time points (temporal quantification). We are hence utilizing spatio-temporal relationships between the meteorological documents in order to compute a combined evaluation of the documents of interest.
This type of intra- and inter-document relationships might look different from those established by hypertext links, and from relationships between parts of a composite document, which are related by its structure. However, all of these relationships can be deployed for retrieval purposes only if corresponding methods for information fusion are available. Fuzzy quantifiers are promising in this respect because they are both human-understandable and sufficiently powerful to handle the required two-dimensional aggregation (data to be aggregated plus weights of relevance). The basic aptitude of fuzzy quantifiers for combining search ratings of a document's parts to its overall evaluation has recently been demonstrated by Bordogna and Pasi (1997).

The span of application even includes traditional bibliographic retrieval, where fuzzy quantifiers provide an alternative analysis of user-assigned term weights, as compared to the so-called "topological" model of Cater and Kraft (1987) and Bordogna et al. (1990); Bordogna and Pasi (1993).

Because of its inconsistency with viewing term-weights as degrees of term-user-relevance, the topological model re-interprets term weights as soft constraints on the term-document-relevance of matching documents (i.e. the users specify 'ideal' term-document-weights).
Fuzzy quantifiers may help to recover the relevance-based view and make it possible to interpret term-weights directly as degrees of term-user-relevance. For example, if W is the fuzzy set of term-user-relevances (i.e. of user-specified term-weights) and if the gradual relevance of the search terms with respect to a given document is expressed by a fuzzy set D, then the quantifying expression **all**(W, D) models weighted conjunction ("all user-relevant terms are document-relevant"), and the quantifying expression **some**(W, D) models weighted disjunction ("at least one user-relevant term is document-relevant"). Genuine fuzzy quantifiers like **many, almost all** etc. offer more subtle ways of aggregation. This novel approach to weighted retrieval will be developed further in a subsequent project (see Glöckner and Knoll (1999c)), which serves to demonstrate its practical utility and to gather empirical results on its benefits.

[8] cf. Silvert (1979); Bloch (1996)

5 Conclusion

We have presented a system architecture suitable for building high-quality multimedia search services for restricted (but in principle arbitrary) topic areas. The architecture is best suited for domains in which there are relatively few document types which possess a large number of document instances. In this case, only a modest number of domain concepts must be implemented, and the implementation of these concepts can be taylored to known characteristics of the document types. These conditions are naturally met in technical or scientific domains: meteorology, satellite imagery, medicine (e.g. radiology), geology and others. The preference for such domains is also typical of other systems which aim at concept-level search, e.g. the SPIRE system presented in Castelli et al. (1998).

Extending HPQS to restricted domains of everyday life such as sports (e.g. soccer games) is certainly a challenge. However, the difficulties of providing high-quality search in such domains are not caused by our architecture, but rather inherent to the complexity of the domain itself. It is our hope that the proposed architecture will provide a framework for building such complex systems.

We have chosen to support a natural language interface in order to remove technical barriers in accessing the system, thus making it useful for a broad public. In addition, support of NL querying is a prerequisite of building future information systems which can be queried by speech. In the medical domain, for example, one can envision a surgeon consulting an intelligent database of X-ray images, patient data etc. by asking questions via microphone.

In this sequel, several major issues have been identified which need to be solved in order to make the techniques of natural language processing useful for broader applications, viz. robustness, lexical coverage, and the problem of disambiguation. The natural language interface NatLink achieves robustness by its incremental parsing strategy and by making use of MESNET graphs, which support semantic representations on different levels of granularity, in dependence on the quality of information available from the analysis.

The problem of lexicon development effort is alleviated by NatLink's support of a hierarchically structured lexicon, to ensure reusability, consistency, small redundancy, and maintainability of information by means of multiple inheritance with defaults, and by provision of an ergonomic lexicographer's workbench.

The choice of disambiguation methods one of the crucial issues because only a high quality of disambiguation can avoid frustration of the users due to wrong interpretations of their queries. NatLink is equipped with innovative disambiguation methods which combine the rule-based and the statistical approaches into a hybrid solution which combines the best of both paradigms: linguistic knowledge can be explicitely coded in the form of rules, and the remaining cases not covered by the encoded knowledge are automatically handled by statistical preferences which are optimised with respect to an annotated corpus. The special cases of PP attachment/PP interpretation have been discussed in order to highlight NatLink's hybrid disambiguation model. In particular, the results of an empirical comparison for German indicate that NatLink typically outperforms existing approaches to PP attachment and multiple PP attachment. With PP interpretation, NatLink has been the first system to gather cross-validated results

for PP interpretation in German. The obtained results of 83.3–92.5% correctness for prepositions with more than one reading are very promising.

Apart from its description of NatLink, the sequel has focused on the retrieval component of the HPQS system. This is because even the best natural language interface would be useless unless there are appropriate retrieval methods for "computing with words", i.e. methods which can provide a linguistically adequate interpretation of natural language queries in the underlying multimedia system.

We have argued that the development of methods which can handle the inherent fuzziness and imprecision of NL queries is crucial to the performance of a retrieval system which can be queried by natural language: an adequate system behaviour can only be achieved if these factors do not give rise to system failure or implausible results. By contrast, an appropriate retrieval model might even profit from these factors by turning them into criteria for result ranking. In order to meet these requirements, we have developed a semantically rich retrieval model based on fuzzy set theory.

Special emphasis had to be put on linguistically motivated methods for information fusion because the "traditional" way of combining search results by applying set-theoretical operations is too weak for NL queries. It is only their use of formal query languages which has permitted traditional retrieval systems to restrict attention to the Boolean connectives **and**, **or** and **not**, because the formal language gives precise control of the set of available aggregation operators. As soon as natural language input is supported, the artificial "barrier" established by the formal query language is removed, and the retrieval module is confronted with the full range of information combining operators of natural language in their full complexity. In the sequel we have presented some results on the most important class of such operators, that of natural language quantifiers. There is an apparent promise of supporting these operators because of their expressive power and ease of understanding. Rather subtle expressions like **almost all** are used by humans with the same ease as the simple Boolean connectives, but they pose extreme difficulties to computational interpretation. We have presented counter-examples from the HPQS domain which show that none of the existing approaches which attempt a numerical interpretation of these quantifiers can provide linguistically adequate models. The fuzzy retrieval component of the HPQS system improves upon previous applications of fuzzy quantifiers because it builds on a novel approach to fuzzy quantification with superior formal properties. Its model of natural language operators is based on an axiomatic characterisation of linguistically adequate interpretations, which avoids implausible results like those presented here from the outset. The HPQS system currently utilizes these operators for evaluating accumulative properties of regions in time or space. However, these operators are truely generic methods for linguistic information fusion, which can also be applied in a broad range of other applications.

Bibliography

Bader, M. (editor) (1995). *Images in Weather Forecasting*. Cambridge, MA: Cambridge University Press.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4:159–219.

Benthem, J. (1983). Determiners and logic. *Linguistics and Philosophy*, 6:447–478.

Benthem, J. (1984). Questions about quantifiers. *Journal of Symbolic Logic*, 49:443–466.

Bimbo, A. D.; Campanai, M.; and Nesi, P. (1993). A three-dimensional iconic environment for image database querying. *IEEE Trans. on Software Eng.*, 19(10):997–1011.

Bimbo, A. D. and Pala, P. (1997). Visual image retrieval by elastic matching of user sketches. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 19(2):121–132.

Biskup, J.; Freitag, J.; Karabulut, Y.; and Sprick, B. (1997). A mediator for multimedia systems. In *Proceedings 3rd International Workshop on Multimedia Information Systems*. Como, Italia.

Bloch, I. (1996). Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(1):52–67.

Bordogna, G.; Carrara, P.; and Pasi, G. (1990). Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27(1):15–26.

Bordogna, G. and Pasi, G. (1993). A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation. *J. of the ASIS*, 44(2):70–82.

Bordogna, G. and Pasi, G. (1997). A fuzzy information retrieval system handling users' preferences on document sections. In Dubois et al. (1997).

Brill, E. and Resnik, P. (1994). A rule-based approach to prepositional phrase attachment disambiguation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pp. 1198–1204.

Bröker, N.; Hahn, U.; and Schacht, S. (1994). Concurrent lexicalized dependency parsing: The ParseTalk model. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*.

Castelli, V.; Bergman, L.; Li, C.-S.; and Smith, J. (1998). Search and progressive information retrieval from distributed image/video databases: the SPIRE project. In Nikolaou and Stephanidis (1998), pp. 157–168.

Catell, R. and Barry, D. (editors) (1997). *The Object Database Standard: ODMG 2.0*. San Francisco, CA: Morgan Kaufmann.

Cater, S. and Kraft, D. (1987). TIRS: A topological information system satisfying the requirements of the Waller-Kraft wish list. In *Proceedings of the tenth annual ACM-SIGIR conference on research and development in information retrieval*, pp. 171–180. New Orleans, Louisianna.

Collins, M. and Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In *Proceedings of the 3rd Workshop on Very Large Corpora (WVLC-3)*. URL http://xxx.lanl.gov/ps/cmp-lg/9506021.

Dubois, D. and Prade, H. (1985). Fuzzy cardinality and the modelling of imprecise quantification. *Fuzzy Sets and Systems*, 16:199–230.

Dubois, D.; Prade, H.; and Yager, R. (editors) (1997). *Fuzzy Information Engineering*. New York: Wiley.

Eimermacher, M. (1988). *Wortorientiertes Parsen*. Ph.D. thesis, TU Berlin, Berlin.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkhani, M.; Hafner, J.; Lee, D.; Petkovic, D.; Steele, D.; and Yanker, P. (1995). Query by image and video content: The QBIC system. *IEEE Computer*, 28(9).

Franz, A. (1996a). *Automatic Ambiguity Resolution in Natural Language Processing*, volume 1171 of *LNAI*. Berlin: Springer.

Franz, A. (1996b). Learning PP attachment from corpus statistics. In Wermter et al. (1996), pp. 188–202.

Fuhr, N. (1995). Probabilistic Datalog - a logic for powerful retrieval methods. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (edited by Fox, E.; Ingwersen, P.; and Fidel, R.), pp. 282–290. New York: ACM.

Glöckner, I. (1997). DFS – an axiomatic approach to fuzzy quantification. TR97-06, Techn. Fakultät, Univ. Bielefeld.

Glöckner, I. (1999). A framework for evaluating approaches to fuzzy quantification. TR99-03, Techn. Fakultät, Univ. Bielefeld.

Glöckner, I. and Knoll, A. (1997). Fuzzy quantifiers for processing natural-language queries in content-based multimedia retrieval systems. TR97-05, Technische Fakultät, Universität Bielefeld.

Glöckner, I. and Knoll, A. (1999a). Application of fuzzy quantifiers in image processing: A case study. In *Proceedings of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems (KES '99)*. Adelaide, Australia. (to appear).

Glöckner, I. and Knoll, A. (1999b). A framework for evaluating fusion operators based on the theory of generalized quantifiers. In *Proceedings of the 1999 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '99)*. Taipei, Taiwan. (to appear).

Glöckner, I. and Knoll, A. (1999c). Konzeption und Implementierung eines kooperativen Assistenten zur Recherche in den Datenbanken der wissenschaftlichen Bibliotheken. In *23. Jahrestagung der Gesellschaft für Klassifikation e.V.: Klassifikation und Informationsverarbeitung zur Jahrtausendwende*.

Glöckner, I. and Knoll, A. (1999d). Natural language navigation in multimedia archives: An integrated approach. In *Proceedings of the Seventh ACM Multimedia Conference (MM '99),*. Orlando, Florida. (to appear).

Glöckner, I. and Knoll, A. (1999e). Query evaluation and information fusion in a retrieval system for multimedia documents. In *Proceedings of the Second International Conference on Information Fusion (Fusion'99)*, pp. 529–536. Sunnyvale, CA.

Glöckner, I.; Knoll, A.; and Wolfram, A. (1998). Data fusion based on fuzzy quantifiers. In *Proc. of EuroFusion98, Int. Data Fusion Conference*, pp. 39–46.

Hartrumpf, S. (1996). *Redundanzarme Lexika durch Vererbung*. Master's thesis, Universität Koblenz-Landau, Koblenz.

Hartrumpf, S. (1997). Partial evaluation for efficient access to inheritance lexicons. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*, pp. 43–50. Tzigov Chark, Bulgaria. URL http://xxx.lanl.gov/abs/cmp-lg/9808013.

Hartrumpf, S. (1999a). Hybrid disambiguation of prepositional phrase attachment and interpretation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, pp. 111–120.

Hartrumpf, S. (1999b). Partial evaluation for efficient access to inheritance lexicons. In *Recent Advances in Natural Language Processing: Selected Papers from RANLP'97* (edited by Mitkov, R. and Nicolov, N.), pp. n–n+11. Amsterdam: John Benjamins. Forthcoming.

Helbig, H. (1986). Syntactic-semantic analysis of natural language by a new word-class controlled functional analysis. *Computers and Artificial Intelligence*, 5(1):53–59.

Helbig, H. (1997a). Der MESNET Primer – Die Darstellungsmittel der Mehrschichtigen Erweiterten Semantischen Netze. Technische Dokumentation, FernUniversität Hagen, Hagen, Germany.

Helbig, H. (1997b). Multilayered extended semantic networks as an adequate formalism for meaning representation and cognitive modelling. In *5th International Cognitive Linguistics Conference*. Amsterdam.

Helbig, H.; Böttger, H.; Ficker, F.; String, P.; and Zänker, F. (1990). The natural language interface NLI-AIDOS. *Journal of New Generation Computing Systems*, pp. 221–246.

Helbig, H.; Gnörlich, C.; and Menke, D. (1997). Realization of a user-friendly access to networked information retrieval systems. In *Proceedings of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, pp. 62–71. Stanford, CA.

Helbig, H. and Hartrumpf, S. (1997). Word class functions for syntactic-semantic analysis. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*, pp. 312–317. Tzigov Chark, Bulgaria.

Helbig, H. and Mertens, A. (1994). Word Agent Based Natural Language Processing. In *Proceedings of the 8th Twente Workshop on Language Technology – Speech and Language Engineering, Twente, 1 and 2 December 1994* (edited by Boves, L. and Nijholt, A.), pp. 65–74. Universiteit Twente, Fakulteit Informatica, Enschede.

Helbig, H. and Schulz, M. (1997). Knowledge representation with MESNET: A multilayered extended semantic network. In *Proceedings of the AAAI Spring Symposium on Ontological Engineering*, pp. 64–72. Stanford, CA.

Hermes, T.; Klauck, C.; Kreyß, J.; and Zhang, J. (1995). Image retrieval for information systems. In *Proc. SPIE's Symp. on Electronic Imaging*. San Jose.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Hutchison, K. and Hardy, K. (1995). Threshold functions for automated cloud analyses of global meteorological satellite imagery. *Int. J. Remote Sensing*, 16(18):3665–3680.

IBM Corp. (1997). http://www.software.ibm.com/data/mediaminer/.

Iyengar, S. and Kashyap, R. (1988). Special section on image databases. *IEEE Trans. on Software Eng.*, 14(5):608–688.

Keenan, E. and Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, 9.

Klavans, J. L. and Resnik, P. (editors) (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Language, Speech, and Communication. Cambridge, Massachusetts: MIT Press.

Knoll, A.; Altenschmidt, C.; Biskup, J.; Blüthgen, H.-M.; Glöckner, I.; Hartrumpf, S.; Helbig, H.; Henning, C.; Karabulut, Y.; Lüling, R.; Monien, B.; Noll, T.; and Sensen, N. (1998a). An integrated approach to semantic evaluation and content-based retrieval of multimedia documents. In Nikolaou and Stephanidis (1998), pp. 409–428.

Knoll, A.; Glöckner, I.; Helbig, H.; and Hartrumpf, S. (1998b). A system for the content-based retrieval of textual and non-textual documents using a natural language interface. Informatik-Bericht 232, FernUniversität Hagen, Hagen, Germany. URL http://ki37.fernuni-hagen.de/papers/ib232.pdf.

Liu, Y. and Kerre, E. (1998). An overview of fuzzy quantifiers. (I). interpretations. *Fuzzy Sets and Sys.*, 95:1–21.

Manber, U. and Wu, S. (1993). GLIMPSE: A tool to search through entire file systems. TR 93-34, Department of Computer Science, University of Arizona, Tucson, Arizona.

Mehl, S.; Langer, H.; and Volk, M. (1998). Statistische Verfahren zur Zuordnung von Präpositionalphrasen. In *Proceedings of the 4th Conference on Natural Language Processing – KONVENS-98* (edited by Schröder, B.; Lenders, W.; Hess, W.; and Portele, T.), number 1 in Computers, Linguistics, and Phonetics between Language and Speech, pp. 97–110. Frankfurt, Germany: Peter Lang.

Merlo, P.; Crocker, M. W.; and Berthouzoz, C. (1997). Attaching multiple prepositional phrases: Generalized backed-off estimation. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, pp. 149–155. Brown University, Providence, Rhode Island.

Nikolaou, C. and Stephanidis, C. (editors) (1998). *Research and Advanced Technology for Digital Libraries: Proceedings of ECDL '98*, LNCS 1513. Springer.

Pinkal, M. (1983). On the limits of lexical meaning. In *Meaning, Use and Interpretation of Language* (edited by Bäuerle, R.; Schwarze, C.; and Stechow, A.), pp. 400–422. Berlin: Walter de Gruyter.

Ralescu, A. (1986). A note on rule representation in expert systems. *Information Sciences*, 38:193–203.

Ralescu, D. (1995). Cardinality, quantifiers, and the aggregation of fuzzy criteria. *Fuzzy Sets and Systems*, 69:355–365.

Ratnaparkhi, A. (1998). Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pp. 1079–1085. URL http://xxx.lanl.gov/ps/cmp-lg/9807011.

Ratnaparkhi, A.; Reynar, J.; and Roukos, S. (1994). A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Human Language Technology Workshop*, pp. 250–255.

Schulz, M. (1999). *Eine Werkbank zur interaktiven Erstellung semantikbasierter Computerlexika*. Ph.D. thesis, FernUniversität Hagen, Hagen, Germany.

Schulz, M. and Helbig, H. (1996). COLEX: Ein Computerlexikon für die automatische Sprachverarbeitung. Informatik-Bericht 210, FernUniversität Hagen, Hagen, Germany.

Sensen, N. (1999). Algorithms for a job scheduling problem within a parallel digital library. In *Proc. of ICCP'99*. (to appear).

Silvert, W. (1979). Symmetric summation: A class of operations on fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:657–659.

Stetina, J. and Nagao, M. (1997). Corpus-based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the 5th Workshop on Very Large Corpora (WVLC-5)*, pp. 66–80.

Virage Inc. (1997). http://www.virage.com/market/vir.html.

Wermter, S.; Riloff, E.; and Scheler, G. (editors) (1996). *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, volume 1040 of *LNAI*. Berlin: Springer.

Yager, R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. on Sys., Man, and Cyb.*, 18(1):183–190.

Yager, R. (1991). Connectives and quantifiers in fuzzy sets. *Fuzzy Sets and Systems*, 40:39–75.

Yager, R. (1993). Counting the number of classes in a fuzzy set. *IEEE Trans. on Sys, Man, and Cyb.*, 23(1):257–264.

Yeh, A. S. and Vilain, M. B. (1998). Some properties of preposition and subordinate conjunction attachments. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, pp. 1436–1442.

Zadeh, L. (1979). A theory of approximate reasoning. In *Mach. Intelligence* (edited by Hayes, J.; Michie, D.; and Mikulich, L.), volume 9, pp. 149–194. New York: Halstead.

Zadeh, L. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers and Math. with Appl.*, 9:149–184.

Zhang, J.; Knoll, A.; and Renners, I. (1999). Efficient learning of non-uniform B-splines for modelling and control. In *International Concerence on Computational Inteligence for Modelling, Control and Automation*, pp. 282–287. Vienna.