Argumentative Zoning:

Information Extraction from Scientific Text

Simone Teufel

PhD
University of Edinburgh
1999

Acknowledgements

Let me tell you, writing a thesis is not always a barrel of laughs—and strange things can happen, too. For example, at the height of my thesis paranoia, I had a recurrent dream in which my cat Amy gave me detailed advice on how to restructure the thesis chapters, which was awfully nice of her. But I also had a lot of human help throughout this time, whether things were going fine or beserk.

Most of all, I want to thank Marc Moens: I could not have had a better or more knowledgable supervisor. He always took time for me, however busy he might have been, reading chapters thoroughly in two days. He both had the calmness of mind to give me lots of freedom in research, and the right judgement to guide me away, tactfully but determinedly, from the occasional catastrophe or other waiting along the way. He was great fun to work with and also became a good friend.

My work has profitted from the interdisciplinary, interactive and enlightened atmosphere at the Human Communication Centre and the Centre for Cognitive Science (which is now called something else). The Language Technology Group was a great place to work in, as my research was grounded in practical applications developed there.

Jean Carletta helped me design the annotation experiment and interpret the results. Her keen eye and sharp mind helped me nip many errors in the bud. I have also enjoyed many exciting discussions with Chris Brew, about statistics, scientific writing and many other things. Katja Markert and Michael Strube read and commented on versions of chapters.

My annotators, Vasilis Karaiskos, Anne Wilson and Anders Bowers, did meticulous work; their critical comments on the task, the guidelines, and their observations about the articles were extremely valuable. Thanks also to the subjects who participated in the smaller annotation experiment.

David McKelvie and Claire Grover from the Language Technology Group expertly helped me with problems to do with XML encoding and transformation, fsgmatch grammar debugging, cascading style sheets and the like. Andrei Mikheev also helped me with statistical and practical problems. I was often thankful for his programs which have proven very useful for my task. Frank Keller and David Sterrat singlehandedly solved the one or two LaTeXproblems I might have had :-).

I was also blessed with the finest crowd of friends in Edinburgh: Frank and

Mirella, Katja, Vasilis, Janet, Claire, Marc, Scott, Elina, Mercè, Karen, Ash, Line, Zelal, Massimo, Roberto, Ira and Jesse. My youngest friends in Edinburgh, Thomas and Catrin, have occupied a special place in my heart. Richard, for reading the damn thing twice, for staying up (remotely, in Berlin!) during the frantic night of last corrections and printing, for sorting out my misbehavin' linux—but even more so for dragging me away from my desk from time to time to lovely places like Gullane beach et al., and for being there.

Last, and most importantly, I want to thank the poor people who are contractually bound not to give up on me. Liebe Eltern, ich danke euch für euren unerschütterlichen Glauben an mich, für eure Engelsgeduld, eure Liebe und Unterstützung, und eure Toleranz in allen Dingen. Ohne euch hätte ich das nie geschafft.

Abstract

We present a new type of analysis for scientific text which we call *Argumentative Zoning*.

We demonstrate that this type of text analysis can be used for generating user-tailored and task-tailored summaries and for performing more informative citation analyses.

We also demonstrate that our type of analysis can be applied to unrestricted text, both automatically and by humans. The corpus we use for the analysis (80 conference papers in computational linguistics) is a difficult test bed; it shows great variation with respect to subdomain, writing style, register and linguistic expression. We present reliability studies which we performed on this corpus and for which we use two unrelated trained annotators.

The definition of our seven categories (argumentative zones) is not specific to the domain, only to the text type; it is based on the typical argumentation to be found in scientific articles. It reflects the attribution of intellectual ownership in scientific articles, expressions of authors' stance towards other work, and typical statements about problem-solving processes.

On the basis of sentential features, we use two statistical models (a Naive Bayesian model and an ngram model operating over sentences) to estimate a sentence's argumentative status, taking the hand-annotated corpus as training material. An alternative, symbolic system uses the features in a rule-based way.

The general working hypothesis of this thesis is that empirical discourse studies can contribute to practical document management problems: the analysis of a significant amount of naturally occurring text is essential for discourse linguistic theories, and the application of a robust discourse and argumentation analysis can make text understanding techniques for practical document management more robust.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Simone Teufel*)

Contents

1 Introduction **13**
- 1.1 Information Foraging in Science 13
- 1.2 Scientific Articles 17
- 1.3 Empirical Natural Language Research 18
- 1.4 Goal and Outline of this Thesis 21

2 Motivation **25**
- 2.1 Manual Abstracting 26
  - 2.1.1 Summary Tailoring 26
  - 2.1.2 Citation Information 31
- 2.2 Automatic Abstracting 35
  - 2.2.1 Text Extraction 37
  - 2.2.2 Fact Extraction 42
- 2.3 A New Approach 48
  - 2.3.1 Design of the Approach 48
  - 2.3.2 Rhetorical Document Profiles (RDPs) 57
  - 2.3.3 RDPs for Tailored Summaries 64
  - 2.3.4 RDPs for Citation Maps 70
- 2.4 Conclusion 73

3 Argumentative Zoning **75**
- 3.1 Fixed Section Structure 77
- 3.2 A Model of Prototypical Scientific Argumentation 83
  - 3.2.1 Argumentative Moves: Swales (1990) 83
  - 3.2.2 Citations and Author Stance 88
  - 3.2.3 Attribution of Intellectual Ownership 93
  - 3.2.4 Statements about Problem-Solving Processes 97
  - 3.2.5 Scientific Meta-Discourse 99

3.2.6 Strategies of Scientific Argumentation 102
3.3 An Annotation Scheme for Argumentative Zones 106
3.4 Argumentative Zones and RDP Slots 114
3.5 Related Work 118
3.6 Conclusion 125

4 A Gold Standard for Argumentative Zoning 129
4.1 Evaluation Strategy 130
4.1.1 Evaluation of Fact Extraction 130
4.1.2 Evaluation of Text Extraction 131
4.1.3 Our Evaluation Strategy for Argumentative Zoning 138
4.2 Evaluation Measures 141
4.3 Reliability Studies 145
4.3.1 Experimental Design 145
4.3.2 Study I 146
4.3.3 Study II 152
4.3.4 Study III 159
4.3.5 Significance of Reliability Results 161
4.4 Post-Analyses 162
4.4.1 Argumentative Structure of Author Abstracts 162
4.4.2 Reduction of Annotation Areas 164
4.5 Conclusion 168

5 Automatic Argumentative Zoning 171
5.1 Overview of Automatic Argumentative Zoning 171
5.2 Correlates of Argumentative Status 174
5.2.1 Traditional Features 178
5.2.2 Meta-Discourse Features 186
5.3 A Prototype System 195
5.3.1 Corpus Encoding 196
5.3.2 Preprocessing 199
5.3.3 Feature Determination 200
5.3.4 Statistical Classifiers 214
5.3.5 Symbolic Rules 219
5.4 Intrinsic Evaluation 220
5.4.1 Naive Bayes Model 220

5.4.2 N-Gram Model 225
5.4.3 Symbolic Rules 227
5.5 Results of System Run on Unseen Material 227
5.6 Conclusion 233

6 Conclusions **237**
6.1 Contribution of the Thesis 238
6.2 Future Work 239
6.2.1 RDP Generation 239
6.2.2 Improving the Prototype 240
6.2.3 Learning Meta-discourse Expressions 244
6.2.4 Redefining the Annotation Task 245
6.2.5 Application to a Different Domain 246

Bibliography **247**

A The Corpus **271**
A.1 Format of Article Encoding 271
A.2 List of Scientific Articles 273

B Example Paper cmp_lg-9408011 **277**
B.1 XML Format 277
B.2 As Published 285
B.3 RDP 294
B.4 RDP Sentence Material 296
B.5 Human Annotation (Annotator A) 298
B.6 Human Annotation (Annotator B) 299
B.7 Agent and Action Recognition 300
B.8 Automatic Annotation (Naive Bayes) 302
B.9 Automatic Annotation (N-Gram) 303

C Annotation Materials **305**
C.1 Study I: Guidelines for Human Annotation of Basic Scheme 305
C.2 Study II: Guidelines for Human Annotation of Full Scheme 312
C.3 Study III: Short Instructions for Human Annotation 329

D Lexical Resources **331**
D.1 Formulaic Patterns 331
D.2 Agent Patterns 339
D.3 Action Lexicon 343
D.4 Concept Lexicon 345

Index of Citations **348**

Chapter 1

Introduction

The topic of this thesis is information management for researchers. Information management is a task that has attracted the attention of researchers in information retrieval and recently also researchers in artificial intelligence and natural language processing. The management of information contained in *scientific articles* poses specific problems. This introduction will set the scene by elaborating what is special about scientific articles. Before we describe the specific goal of this thesis, we will introduce the data we work with: a corpus of "real-life" computational linguistics conference articles. We will also discuss why we find this topic interesting, both from a research perspective as well as from a practical one.

This discussion will result in our general hypotheses for this work. We will argue for the application of empirical discourse studies when tackling document management problems. We believe that the argumentative analysis of naturally occurring text can provide subject-matter independent information which can fulfil many searchers' information needs, particularly the needs of less experienced searchers.

1.1. Information Foraging in Science

In today's fast moving academic world, new conferences, journals and other publications are springing into existence and are expanding the already huge repository of scientific knowledge at an alarming rate. Cleverdon (1984) estimates an annual output of 400,000 papers from the most important journals covering the natural sciences and technology. Kircz (1998) states that Physics Abstracts, the major bibliographic abstracting service in physics and the manufacturer of the INSPEC database, indexed 174,000 items in one year alone (1996), of which about 146,500 are journal articles. However, these already impressive numbers exclude less important journals, workshop proceedings, conference papers and non-English material. Indeed, the growth rate is probably exponential—Maron and Kuhns (1960) estimated that the indexed scientific material doubles in volume every 12 years.

The masses of information the researcher is exposed to make it hard for her to find the needle in the haystack as it is impossible to skim-read even a portion of the potentially relevant material. The information access and search problem is particularly acute for researchers in interdisciplinary subject areas like computational linguistics or cognitive science, as they must in principle be aware of articles in a whole range of neighbouring fields, such as computer science, theoretical linguistics, psychology, philosophy and formal logic.

Apart from keeping abreast of developments in scientific fields in general, more practical requirements emerge when researchers who are experienced in one scientific field start getting interested in a *new* scientific field, in which they have no prior knowledge. Their information needs have suddenly changed: Kircz (1991) states that such readers seek understanding instead of a firm, formal answer. The exact information need is not known beforehand; the questions they pose are not precise (Kircz' example is the question *"what are they doing in high-temperature super-conductivity?"* (p. 357)). Belkin (1980) refers to their situation as an "anomalous knowledge state". We think that researchers in a new field initially need answers to the following questions:

What are the main problems and main approaches? Knowledge of a number of important concepts in the field needs to be acquired: the current problems and the standard methodologies in the field. For the main approaches, the researcher needs to know their strengths and weaknesses. The searcher also needs to gain an overview of the evaluation methodology and typical numerical results in the field.

Which researchers and groups are connected with which concepts? Researchers' names—and the institutions where they work—must be associated with seminal approaches and seminal papers. The searcher must determine *schools of thought*: clusters of people working together, sharing premises and building on each others work.

If researchers read a paper in a new field, they are particularly interested in the general approaches described, the relation to other work, and its conclusions, instead of specialist details (Kircz, 1991). Oddy et al. (1992) and Shum (1998) argue that what such readers particularly need is an embedding of the particular piece of work within a broader context and in relation to other works.

The preferred information source at that stage of knowledge is an experienced colleague. Another standard technique for gaining a deeper overview of a field is to find a recent review article, to follow up the bibliographic links and to read however many of those papers one's time permits.

But sometimes neither of these useful aids is available, and a full-blown bibliographic search using an electronic document retrieval system is necessary, e.g. BIDS, FirstSearch or MEDLINE. This is typically done by a keyword search, where the keywords can be combined with Boolean operators.

In most commercial bibliographic data bases, keyword search is still performed on *document surrogates*, rather than on the full text of the document, as the full text is not always available in electronic form. Typical document surrogates used in document retrieval environments are bibliographic information (i.e. title, authors, date of publication, journal name), a list of index terms, or a human-written summary. The assumption is that these document surrogates capture an important aspect of the meaning of the document, i.e. that they are able to give the searcher a characterization of the contents of the paper, and that they can thus be used as a search ground. Mathematically sophisticated matching procedures between the document surrogates and the user's query measure how appropriate the document is for a certain query (*query-document similarity*). Document surrogates are also used to present the search result to the searcher, typically as an unordered list. The user can then perform *relevance assessment* on the basis of the document surrogates, i.e., she can filter out the obviously irrelevant documents from the search results.

There is a wide range of empirical studies about users of online data bases (Bates, 1998; Borgman, 1996; Fidel, 1985, 1991; Saracevic et al., 1988; Ellis, 1992; Ingwersen, 1996). These studies look at many different factors like searching experience, task training, educational level, type of search questions and user goals. The few of these studies which include inexperienced users conclude that the state of the art in document retrieval systems puts less experienced users at a disadvantage: those who have less well-defined queries and information needs (Clove and Walsh, 1988).

As they know neither the basic concepts nor the terminology of the new field, such searchers cannot possibly do well on keyword searches. The search terms they choose are often too unspecific and produce too many hits (Ellis, 1989a,b), hits where the term has another meaning, or no hits at all. As most search engines for bibliographic search rely on Boolean search and return the search results as an unranked list, they are at risk of getting lost in the returned list of document surrogates. Kircz (1991) calls this phenomenon the “frustrating circularity of the Boolean search process”: clean, relevant information can only be retrieved from a data base if the searcher already knows what she is looking for.

Inexperienced searchers also have problems with the relevance decision itself. They cannot be sure that the retrieved articles are relevant to them or if they contain so-called *false negatives*. On the other side, and even more frustratingly, they must suspect that a myriad of relevant articles are in the database which their search has *not* found (*false positives*). (False negatives and positives are a normal phenomenon in free-text search; they are caused by polysemy and synonymy and by more complex features of unrestricted language.) To have access to high-quality document surrogates would be very important to the searchers—good abstracts are essential, as these are often the first detailed indication of the document’s contents that they see. Titles alone are typically not informative enough for them.

However, even with imperfect search there is typically a convergence towards a few seminal papers which are frequently cited—even if the searcher was unlucky enough to start the search with peripheral, controversial or weak papers (along with the outright irrelevant ones). However, this is a more or less random process which might require a long time.

There are many ways in which this situation could be ameliorated, e.g. by better search methods or by better presentation of the search results. Best match (i.e. ranking) search algorithms rely on the intuition that it is crucial to get the right papers to the user in the right order, e.g. Salton’s (1971) SMART system, or Robertson et al.’s (1993) OKAPI system.

The retrieved items can also be displayed by *document–document* similarity rather than by *query–document* similarity, e.g. VIBE (Olsen et al., 1993), Scatter/Gather (Hearst and Pedersen, 1996), Vineta (Krohn, 1995), Bead (Chalmers and Chitson, 1992), TileBars (Hearst, 1995) and Envision (Nowell et al., 1996).

In this thesis we will choose a different route: in the line of automatic abstracting approaches, we aim to improve the document surrogates returned to the searcher. We believe that better document surrogates will not only support the searchers in their relevance decision but it should also improve search itself. We believe that it is particularly important to design document surrogates which represent information needs that are typical for new searchers. In order to generate such document surrogates, the *right* kind of information must be extracted from the articles. This thought is one of the starting points for the present thesis.

1.2. Scientific Articles

One of the reasons why we chose to work with scientific articles is the practical value of better document retrieval environments for scientists. Scientific research articles are the main source of current leading-edge information for researchers, rather than text books or other sources of scientific information. In a library setting, there is a realistic demand for better summaries, or better document surrogates in general, cf. the recent interest in digital libraries.

The other motivation is more theoretical. Scientific papers are different from other text types with respect to their overall structure, an aspect we are particularly interested in. For a start, they are not organized in a time-linear manner. Assumptions about time linearity might help with the processing and summarization of simple narrative and newspaper text. Even though scientific articles are reports of intellectual work which was conducted within a certain time frame, their presentation follows the chronological order only in exceptional cases. Instead, the article structure usually mirrors the internal problem space and the scientific argumentation. The clear communicative function of scientific articles and the text-type specific expectations based on this function can provide a possible handle for subject matter-inspecific information extraction from such articles.

The writing style in scientific articles shows a considerable level of variation. Some articles are overtly argumentative, arguing against another author's views; others present empirical work such as a linguistic survey or corpus study in a more objective manner; some describe practical work like an implementation for a given problem. In interdisciplinary fields, articles might combine research methodologies from more than one discipline, e.g. a computational simulation of human behaviour originally observed in a psychological experiment. The linguistic expressions occurring in the articles mirror this variety.

Scientific articles are also biased; they describe the author's work from her own viewpoint. This bias is an integral part of the communicative function of scientific articles: they were written to convince the reader of the validity of a given research. The texts thus typically contain explicit markup of this rhetorical information (*meta-discourse*). In contrast, news stories have a supposedly neutral news anchor, and narrations are often told by an omniscient, neutral narrator. We are interested in the author's bias and aim to exploit it for our task.

Scientific text is harder to analyze than the texts typically used in discourse linguistic approaches. The reason for this is that it is not trivial *which* kind of document structure underlies scientific articles. Grosz and Sidner (1986) analyze apprentice–experts dialogues with an obvious task-structure; Iwanska's (1985) procedural texts are similarly structured. Other texts used for discourse analysis are short and well-edited; cf. Marcu's (1997b) popular science texts. Our texts, in contrast, are more difficult.

We chose *computational linguistics* (CL) as a domain for a number of reasons. One reason is that it is a domain we are familiar with. This makes an intermediate evaluation of our work possible without requiring the judgement of external subject experts. The more theoretically interesting reason is that computational linguistics is a *heterogeneous* domain due to its multidisciplinarity: the papers in our collection cover a wide range of subject matters, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. This results in large differences in document structure and forces us to choose a more domain independent approach to document structure. In sum, our collection is an exciting and challenging test bed for discourse analysis.

1.3. Empirical Natural Language Research

Corpus-based or empirical natural language research is the study of language based on examples of real life language use. It is a general methodology which has come back into fashion recently, and which is now applied in several tasks in theoretical linguistics and natural language processing, e.g. lexicography, syntax and lexical semantics (Manning and Schütze, 1999). The general idea is that a linguist's or system developer's introspection alone cannot predict the unexpected turns of real language use. Rather than dealing with invented or artificially simplified examples, a large sample of naturally occurring language should be used instead. Empirical linguists aim to describe as much of the data as possible, but accept the fact that it is not normally the case that 100% of the data can be accounted for.

It is generally accepted that large corpora are a reliable source of frequency-based data. Additionally, a corpus is a more powerful scientific methodology than introspection as it is open to verification of results (Leech, 1992).

We subscribe to this general methodology: if one is planning to develop a practical system for unrestricted and thus unpredictable text, it is indispensable to base the design of this system on some kind of corpus analysis.

Whereas the Message Understanding Conferences (e.g. MUC-7 1998) have provided several corpora of newspaper articles with answer keys which are readily used in the field (cf. section 2.2.2), researchers wanting to work on scientific articles are at a disadvantage. At the time when research on this thesis started, there was no corpus of scientific articles available, so we collected our own corpus. It was also generally agreed at the AAAI Spring Symposium 1998 for Intelligent Text Summarization (Radev and Hovy, 1998) that there is a real lack of corpora of scientific articles. A version of our corpus is now distributed by TIPSTER as part of the SUMMAC program (Tipster SUMMAC, 1999).

We are interested in naturally occurring, unrestricted text, and we wanted to choose data which is as representative of the field as possible. We chose the Computation and Language Archive (CMP_LG, 1994) as our source, which is part of the CoRR (Computing Research Repository), a large preprint archive.

The idea of a preprint archive is the rapid dissemination of work: researchers can make their results available to the community early, e.g. before the conference where the paper is presented. The preprint version can later be replaced with the published version. Preprint archives, if widely used within a community, are perhaps the best way to track new work, although there is not necessarily a guarantee that the work is peer reviewed.

Between its beginnings in April 1994 and the submission date of this thesis, 968 articles have been put into the CMP_LG archive. The archive seems to be commonly used in the field: for example, researchers in computational linguistics use CMP_LG numbers as a standard way of identifying their papers.

We collected *all* documents from CMP_LG deposited between 04/94 and 05/96 which fulfilled our selection criteria, e.g. they had to have an abstract and be available in LaTeX. All these criteria are formal and not content-based; they are described in full in sections 5.3.1 and 5.3.2, where details about the corpus collection work are given.

One of our selection criteria concerns where the papers were published. We chose what we perceived to be the most influential conferences in CL, namely the *Annual Meeting of the Association for Computational Linguistics* (ACL), the *Meeting of the European Chapter of the Association for Computational Linguistics* (EACL), the *Conference on Applied Natural Language Processing* (ANLP) and the *International Conference on Computational Linguistics* (COLING). As a result, we know that all our papers had been peer reviewed. Restriction to these conferences does not introduce a bias, as CL is a field with few journals, where conferences are very important, and as the chosen conferences are the most influential ones. We also included papers presented in the student sessions, and those published in the proceedings of ACL-sponsored or EACL-sponsored workshops.

The deposition of articles on a preprint archive is voluntary and not systematic; some researchers might choose not to contribute their articles at all, whereas others might deposit an unrepresentatively high number of their articles. It is therefore difficult to claim that our corpus is representative of the *field* of CL as such. However, due to the unbiased sampling procedure, our collection should be reasonably representative of computational linguistics conference articles published in the given time frame and deposited on the CMP_LG archive: there is no reason to believe that new articles which would fulfill our selection criteria should be systematically different from the articles in our collection.

80 papers passed our selection criteria. They constitute the final, closely inspected corpus used in this thesis; details of the corpus are listed in appendix A.2. Roughly, the largest part of articles (about 45%) describe implementational work, 25% describe theoretical-linguistic work, 20% experimental work (corpus studies or psycholinguistic experiments) and 10% report evaluation (i.e., no completely new method is introduced in these articles; instead, already known systems or theories are compared and evaluatively measured).

Following from the fact that we are using unrestricted, naturally occurring text coming from a prepring archive, our texts display large variability in writing style. Some articles in our collection which do not use fully grammatical English; typing errors abound, and the register varies between formal and extremely informal, as the following two sentences illustrate:

> Formal:
> *While these techniques can yield significant improvements in performance, the generality of unification-based grammar formalisms means that there are still cases where expensive processing is unavoidable.* (S-7, 9502021)

> Informal:
> *This paper represents a step toward getting as much leverage as possible out of work within that paradigm, and then using it to help determine relationships among word senses, which is really where the action is.* (S-158, 9511006)

The corpus contains 333,634 word tokens. Even though this is much smaller than the large scale corpora typically used in corpus-based NLP (natural language processing), it still provides an unbiased resource describing a substantial amount of scientific text in computational linguistics.

For comparative purposes, we also had access to two other corpora: a corpus of agriculture, from Chris Paice's group at the Computer Science department of the University of Lancaster, and a corpus of papers in cardiology, from Prof. Kathleen McKeown's group at the Computer Science Department of Columbia University, NYC. In some cases, we will compare properties of our texts to texts from these corpora.

1.4. Goal and Outline of this Thesis

This thesis aims to contribute towards the automatic generation of document surrogates in the framework of a document retrieval environment for scientific articles. The practical topic of this thesis is how document surrogates can help researchers in their scientific information foraging activities, particularly those researchers who are new in a given field.

The thesis is structured as follows: The next chapter will define the goal in more detail, after a look at summaries in today's document retrieval environments. It will show that traditional human-written summaries are not flexible toward user expertise and task requirements, which is particularly a problem for novice researchers in a field. We argue that document surrogates should capture similarities and differences between related articles, which summaries typically do not. Current methods for automatic abstracting, on the other hand, create summaries which are either too generic, containing too little information to adequately characterize the document, or too inflexible towards unexpected material in the text. To ameliorate these problems, a new document surrogate is introduced: the Rhetorical Document Profile (RDP). It encodes typical information needs of new readers, e.g. global level information like which SOLUTION was introduced in the article, or what the GOAL of the article was. We will argue that RDPs are useful for practical document retrieval applications: flexible summaries can be generated from them, and types of connections between articles can be expressed in a construct called a *citation map*. The rest of this thesis will explore the possibility of creating RDPs automatically by a process of robust text analysis and extraction.

Chapter 3 introduces a new document analysis called *Argumentative Zoning*. Argumentative Zoning concentrates on global discourse information: the rhetorical status of a sentence in relation to the discourse act of the overall paper. It turns out that some of these rhetorical states coincide with the information needs introduced in chapter 2; thus, this chapter also gives a justification for RDPs. Argumentative Zoning is independent of writing style, subject matter, and, to a certain degree, subdomain, but relies on text type specific expectations (communicative acts). Section 3.2 introduces our model of prototypical scientific argumentation. This model is operationalized in section 3.3 by introducing seven different information categories or argumentative zones.

Chapter 4 discusses our evaluation strategy for the new task of Argumentative Zoning, in view of similar tasks (fact extraction, text extraction and dialogue coding tasks). The annotation scheme developed in chapter 3 will be empirically validated with respect to human performance, i.e. we will measure to which degree human judgements of argumentative zones agree. This annotation experiment provides us with quantitative data about the reliability of the scheme, and it also gives us training material for our prototype implementation of Argumentative Zoning.

Chapter 5 documents an experiment in automatic Argumentative Zoning. First, we will describe a pool of sentential features which correlate with the sentence's rhetorical status. Then, we will describe the implementation of a prototype system for automatic annotation: the automatic determination of these features, the statistical classifiers used, and a rule-based alternative implementation. We will then present the results of an intrinsic evaluation of our system.

The conclusions will bring us back to the main working hypothesis of the thesis: that empirical discourse studies can contribute to practical document management problems. In this thesis, we use practical discourse studies (in our case, centered around argumentative zones) to help identify the kind of information in scientific texts which are crucial for searchers' information needs. We experimentally show that humans can be trained to perform Argumentative Zoning consistently, and that this behaviour can be simulated by an algorithm; we consider this as a proof of concept for RDPs and for Argumentative Zones.

In the course of the thesis, the following research questions will be addressed:

- *Discourse linguistics:* Is it possible to analyze the document structure of scientific articles in a subject matter-independent way? At which abstraction level should such an analysis define its units and relations? What are the linguistic signals of this structure?

- *Experimental psychology:* To which extent do humans share intuitions about information and document structure in scientific papers? Can people be trained to apply a fixed annotation scheme for the analysis? In which aspects do the humans' annotation differ and agree most?

- *Computational linguistics and artificial intelligence:* Can we identify algorithmically determinable signals of argumentation and document style in unrestricted text? Which of those can be used for system building and evaluation? How much "understanding" would such a system need to produce acceptable document characterizations?

Chapter 2

Motivation

In this chapter, we will define the goal of this thesis in more detail. We will start with a discussion of the most prominent document surrogates—summaries—and the state of the art in producing them, both manually and automatically.

In section 2.1 we focus on manual summarization. We argue that the current practice of abstracting is undergoing a big change because more and more scientific research text is available in electronic form. The high-quality human-written summaries, deeply rooted in the paper-based publishing world, cannot offer the flexibility towards task and user expertise that becomes more and more of a necessity. We will argue that one of the problems of current summaries is that they do not take connections between articles into account.

Section 2.2 will start with an overview of two current automatic summarization methods: text extraction and fact extraction methods. Both have advantages and drawbacks: inflexibility in the case of fact extraction method, the lack of context-sensitivity in the case of text extraction.

In section 2.3 we suggest an approach which synthesizes text and fact extraction methods by attaching global-level rhetorical information to extracted sentences. This results in *Rhetorical Document Profiles* (RDPs). We argue that RDPs combine the best of both worlds from fact extraction and text extraction methods, and that they have definite advantages in a document retrieval environment. We then show how the information contained in them could be used to generate tailored summaries and annotated *citation maps*.

2.1. Manual Abstracting

Humans are well-known to be good summarizers (Kintsch and van Dijk, 1978; Sherrard, 1985; Brown and Day, 1983), and summaries written by well-trained information specialists are of particularly high quality (Lancaster, 1998; Cremmins, 1996). However, as we will see, this is not enough to immediately solve all of the researchers' search problems introduced in the previous chapter.

2.1.1. Summary Tailoring

Information services (secondary publishers) like the *Institute for Science Information, Inc.* or *Chemical Abstracts Service* specialize in information management for scientists. In order to keep researchers informed of publications in their area of interest, these companies publish, amongst other things, journals with summaries of research material.

Such information services have made a huge investment in the production and dissemination of summaries. They employ information specialists (professional abstractors/indexers), highly qualified professionals who have been trained in the art of summarizing and indexing articles and books.

Professional summaries are written according to agreed guidelines and recommendations (McGirr, 1973; Borko and Chatman, 1963; ANSI, 1979; ISO, 1976). The guidelines are concerned with the informativeness and readability of the human-written summaries; they try to make sure that they are general, long-lived and high-quality accounts of the information contained in a scientific article. For example, the guidelines give a certain maximum and minimum number of words to be used in a summary. They recommend that summaries should be aimed at a particular kind of reader, a semi-expert: somebody who knows enough about the field to understand basic methodology and general goals but who would not understand all specialized detail. Also, the summaries are supposed to be self-contained (Lancaster, 1998, p. 108): the reader should be able to grasp the main goals and achievements of the full article without needing the source text for clarification.

In the literature on human summarization we find very little about the tasks that users are assumed to perform with the summaries. The only mention of summary use we find is at an abstract level (e.g. in Lancaster 1998):

1. Summaries can be used as substitutes for the whole document. If researchers

want to be kept *aware* of new publications in a field, it is often enough for them to read summaries in abstract journal (alerting function), instead of reading the full article.

2. Another example of substitutive use of summaries is when they are used to *refresh* a reader's memory of a previously read article.

3. Another situation is the use of summaries in parallel with the full text, e.g. when *previewing* of the structure of the source document. Here, the summary serves as orientation about the structure of a document that has already been chosen, similar to a table of contents.

4. Rarely, summaries are used for reasons having nothing to do with the original text. For example, when users need to decide if they have chosen the *right data base* for a search, they can looked at a random summary of that data base for mere seconds.

5. The most typical use of summaries in a document retrieval environment is for relevance decision, i.e., to judge whether or not the corresponding, as yet unknown, full article is relevant to searchers' current information need (Cremmins, 1996; Rowley, 1982). During this step, the reader might also recognize papers she has read before. The relevance decision process will determine a set of probably relevant papers, which can then be looked up in the library, requested in full from the author or ordered as paper copies. A similar use is the decision of whether or not the searcher has read an article already.

Typically, there is only one version of the summary. The only generally accepted dimensions of summary variance in the literature are compression (i.e. length of summary in comparison to the full text) and the distinction between *indicative* and *informative* summaries. Indicative summaries contain an indication about the topic of the text (i.e., they contain purpose, scope or methodology), whereas informative summaries also name the main findings and conclusions of the text (Rowley, 1982; Cremmins, 1996; Lancaster, 1998; Michaelson, 1980; Maizell et al., 1971). Indicative summaries are of use for relevance decision and all functions which assume that the full text is either available, or that an indication of the general contents is enough for the researcher. Informative summaries, on the other hand, are autonomous texts which can be used as full text substitutes.

Consider the following examples from Lancaster (1998, p. 95):

Indicative Summary:

> *Telephone interviews were conducted in 1985 with 655 Americans sampled probabilistically. Opinions are expressed on whether: (1) the establishment of a Palestinian state is essential for peace in the region; (2) U.S. aid to Israel and to Egypt should be reduced; (3) the U.S. should (a) participate in a peace conference that includes the PLO, (b) favor neither Israel nor the Arab nations, (c) maintain friendly relations with both. Respondents indicated whether or not they had sufficient information concerning various national groups in the region.*

Informative Summary:

> *Telephone interviews conducted in 1985 with 655 Americans, sampled probabilistically, brought these results: most (54–56%) think U.S. aid to Israel and Egypt should be reduced; most (65%) favor U.S. participation in a peace conference that includes the PLO; more than 80% consider it important that the U.S. should maintain friendly relations with both Israel and the Arab Countries; 70% believe that the U.S. should favor neither side; most (55%) think that the establishment of a Palestinian state is essential to peace in the region. The Israelis are the best known of the national groups and the Syrians the least known. The Arab-Israeli situation is second only to the conflict in Central America among the most serious international problems faced by the U.S.*

There is disagreement which type of abstract is easier to write. Rowley (1982) argues that indicative abstracts are more difficult to write, and (Manning, 1990) claims the opposite. Most authors distinguish the so-called *informative-indicative* summary, where some results are given (as would be in an informative summary), whereas other parts of the paper are treated only indicatively. Rowley (1982) states that this kind of summary is most commonly used nowadays; Lancaster (1998) (who does not recognize informative-indicative summaries) states that informative summaries are less common than indicative ones.

Informative summaries are further divided into purpose-oriented and findings-oriented summaries, which differ in the order of the information presented (Cremmins, 1996; ANSI, 1979). Findings-oriented summaries present findings (results and conclusions) first. The following examples from Cremmins (1996, p. 109) illustrate that the difference between them is not great.

Purpose-oriented indicative-informative summary:

> *Suggestibility was measured under indirect, auto-, hetero-, and conflicting forms of suggestion by using the Body Sway Test. Healthy and ill students and patients, with and without autogenic training, were tested. Equally strong effects occurred under all four forms of suggestion. Autogenic training affected positive behavior on the test in both healthy and ill students. Negative behavior in this test occurred when autogenic training was lacking. The behavior of female patients was more positive than that of males under conflicting suggestions.*

Findings-oriented indicative-informative summary:

> *Equally strong effects of suggestion occurred under indirect, auto-, hetero-, and conflicting forms when the Body Sway Test was given to healthy and ill students and patients, with and without autogenic training. The training affected positive behavior on the test in both healthy and ill students. Negative behavior in this test occurred when autogenic training was lacking. The behavior of female patients was more positive than that of males under conflicting suggestions.*

Even though Cremmins does not say so explicitly, it seems likely that the two types of summaries support (slightly) different kinds of tasks. For example, the findings-oriented summary might be more useful to a medical researcher trying to spot the kinds of experimental *results* she would need in support of an argument of her own. The difference in order seems to imply a model of summary use in which users sequentially read the summary from the start and stop reading when they have found what they need for their relevance decision (Borko and Bernier, 1975, p. 69). However, we found no empirical studies in the literature which focus on summary reading strategies or which measure the appropriateness of different kinds of summaries for a certain task. In sum, the assumptions in the literature about user tasks are minimal and do little more than support two uses of summaries: a) as texts that give an indication of the contents and b) as autonomous texts.

Another point is the question how to determine what is relevant for a given user at a given time. There are a myriad of reasons why a user would classify a given document as relevant at a given point in time during relevance decision (Rees, 1966). A vast experimental and theoretical literature in information science has been concerned with the slippery concept of relevance (Saracevic, 1975; Schamber et al., 1990). In principle, it is undisputed that the large-scale context influencing the interpretation of a text and the relative importance of a part of the text depends on and comprises the writer and reader of the text and their background, goals, and viewpoints. Even to the same reader, at different points in time, different aspects of the same text might be relevant. Spärck Jones (1990) describes the general problem by saying that pertinence is situational to a unique occasion.

It is hard to argue with Lancaster (1998) when he states that "the abstractor should [. . .] omit other information that readers would be likely to know or that may not be of direct interest to them." (p. 107)—the difficult part is to guess *which type of information* different groups of readers are likely to know. The *informedness of the intended audience* is one of the central points in user tailoring known from text generation (Spärck Jones, 1988; Paris, 1988, 1994). The summarizing industry, however, does not envisage summaries which are responsive to level of expertise of the reader. Though the concept of *subject slanting* (i.e., tailoring the summary to the anticipated interest of its users) is quite common when summaries are produced for the internal use of one organization, rather little slanting takes place in general information services (Herner, 1959).

Kircz (1991) distinguishes between uninformed, partially informed and informed readers. He argues that the level of subject knowledge influences which information readers draw from scientific articles. Uninformed readers read introductions and conclusions, and also overview figures/graphs if present, and the list of references. Partially informed readers read papers particularly for the general approaches described, the relation to other work, and the conclusions. Informed readers, in contrast, can use their scientific background knowledge in a field to find their way in the literature quickly. They typically scan articles fast; only the core of information is read, e.g. the numerical results. As traditional summaries are geared towards partially informed readers, they are therefore often too terse for uninformed readers, and too verbose for informed readers. This poses more of a problem for the uninformed than for the informed reader.

It is important to see that the inflexibility of traditional summaries is rooted in the function of summaries in the paper-based world of publications which we just described. Recently, due to the omni-presence of the world wide web and electronic journals, more and more papers are available in electronic form—it can be expected for the near future that most bibliographic document retrieval environments will provide researchers with electronic versions of the paper during search time. This development has strong influence on what the most appropriate document surrogate for the search task should look like.

Firstly, and rather obviously, the fact that the full paper is available in electronic form is a necessary precondition for realistic automatic summarization. In the early era of summarization, research was restricted by data problems, and articles had to be manually encoded and typed. Now, the manipulability of electronic text makes it possible to summarize millions of papers—different summaries of one paper can be created on the fly, and it is theoretically possible to be flexible towards length, end task and user expertise.

But electronic texts also pose new challenges, as studies of readers in electronic environments show (Dillon, 1992; Levy, 1997; Adler et al., 1998; O'Hara et al., 1998). Kircz (1998) criticizes the fact that new electronic publishing technology has mostly been used to echo the old style of paper-articles in the new medium, rather than employing new functionality. Other work concentrates on reading strategies. For example, on-line browsers like Netscape or Internet Explorer, and previewers like Ghostview or Adobe Acrobat can display the articles directly on-screen, but they cannot yet simulate the physical properties of paper. O'Hara and Sellen (1997) found that this disrupts typical reading strategies of scientists, e.g. the so-called *non-linear* reading (Samuels et al., 1987; Dillon et al., 1989). The non-linear reader jumps in a seemingly arbitrary fashion from the conclusion to the table of contents and scans the section headers and captions, in order to get an ad-hoc idea of the structure of the text. This strategy serves to efficiently build a model of the text's structure as well as to extract the main concepts of the paper, and is a typical reading behaviour for scientists (Pinelli et al., 1984; Bazerman, 1988).

But even though today's browsers might give a suboptimal representation of the article, new, intelligent display mechanisms could exploit and thus compensate for some of the functions of the material paper (O'Hara and Sellen, 1997). One way in which new functionality can help readers in an electronic environment is the support of citation indexes as an additional search strategy, which will be treated in the next section.

2.1.2. Citation Information

There are information search tasks which are specific to research, tasks concerned with *connections* between research outputs (Oddy et al., 1992). Shum (1998) stresses that researchers, a community which is constantly contesting claims, need information about scientific relationships:

> [...] *relationships* are critical for researchers, who invest a lot of energy in articulating and debating different claims about the significance of conceptual structures. (Shum 1998, p. 19, his emphasis)

Such information results in the knowledge-rich cognitive net of information which Bazerman (1985) describes for physicists and Charney (1993) for evolutionists. Experienced researchers know the important names in the field; they know institutions and their specialities and preferred methodologies; they know schools of thought and how they interrelate. These information nets are acquired over time, by reading, through research, at conferences, and by discussions with colleagues.

However, in the course of conducting a *new* piece of research a researcher is likely to come up with immediate questions for which the background knowledge provides no answers. These pressing questions often result in a document retrieval search:

Supportive Data: During the writing of a paper, the researcher might look for support in the literature for a certain claim she needs as a step in the argumentation. She might first want to check if the claim has been previously stated in print; if this is the case, it is necessary to respect that paper's prior claim of intellectual ownership by citing the given paper. Another task is to find out if the given paper is the original citation for the idea, or if that work continues somebody else's work. In interdisciplinary fields, one might need to include specific evidence coming from a particular neighbouring field, e.g. validation of the claim in the form of experimental psychological results.

Differences and contrasts: The researcher might want to check if there are published results that are contradictory to her own. She might also want to find out if there are competitors to her claim, i.e. rival approaches (approaches with the same goals, but a different methodology). Another question might emerge if she has identified a weakness of some other work—she might want to find out if that work has been criticized by somebody else before, and if so, what exactly constituted the prior criticism.

Updates of old research articles: It sometimes happens that a researcher finds an article which contains the right information (e.g. a particular scientific fact or claim needed for her current work), but which happens to have been published a long time ago. It is considered bad practice to cite the old paper without stating what happened in the meantime with respect to the scientific claim. Shum (1998) mentions the following question as pressing for scientists: *"What impact did certain evidence have?"* More recent articles need to be located which either still maintain the same claims (maybe with additional evidence), or contribute counter-evidence. If the original article is a dated *review* article, a special case of this information need applies: *each* cited article needs to be traced forward in time to some more recent research.

Information about the relatedness of scientific articles is available from citation indexes, e.g. the Institute for Scientific Information (ISI)'s multidisciplinary citation indexes (ISI, 1999). Such indexes cover only a small range of journals, which is justified by the fact that a relative small number of journals account for the bulk of significant scientific results (Garfield, 1996). Traditionally, citation indexes are used for bibliometric studies, i.e., to measure the quality and academic impact that a piece of academic work or a journal has (Garfield, 1979)—an approach which has disadvantages as well as advantages (cf. section 3.2.2). In the context of our task, and apart from impact assessment, citation links can be used in two ways:

- Citation links can provide an alternative way of accessing information in the data base.
- Similarities between articles can be determined by their citation behaviour.

Work on article clustering by citations includes bibliographic coupling (Kessler, 1963) (if two articles have similar bibliographies then they must share a topic) and co-citations Small (1973) (if two papers often occur together in other article's bibliographies then they must share a topic). There is an analogy with research on the topology of the world wide web (Kleinberg, 1998), where *authorities* (often-referred-to, seminal pages) and *hubs* (clusters of pages which list many authorities) are identified.

Citation links can also be used for information access. ISI BIDS, for example, allows users to list document surrogates of all articles citing a given one, and many on-line proceedings are internally citation-indexed (SIGMOD, 1999)—articles cited in the paper can be reached directly, but there is also a listing of all articles citing the given article *later*. Recently, tools for citation manipulation with even higher functionality have emerged. The new citation visualization tool CiteSeer, which is part of NEC's digital library ResearchIndex initiative (Giles et al., 1998) performs *Autonomous Citation Indexing*: a citation index is automatically built from all papers available to CiteSeer. References in running text are automatically determined, and the reference list is parsed. Citation forms appearing in slightly different shape in other sources are mapped onto each other. CiteSeer displays the context in which a given citation occurs in running text by showing the sentence containing the physical reference along with snippets of keywords, headlines and adjacent sentences in an extract-style. The following example citation is taken from (Giles et al., 1998, p. 94); it shows a reference to the paper "*Maximum likelihood from incomplete data via the EM algorithm*", published by Dempster et al. in 1977. The following segment has to be read in order to determine how the two papers relate to each other:

> *... other variant algorithms are also seen to be possible. Some key words: EM algorithm, incremental algorithm, free energy, mixtures Submitted to Biometrika 1 Introduction The Expectation-Maximization (EM) algorithm finds maximum likelihood parameter estimates in problems where some variables were unobserved.* **Its widespread applicability was first discussed by Dempster, Laird and Rubin (1977).** *The EM algorithm estimates the parameters iteratively, starting from some initial guesses. Each iteration consists of an Expectation (E) step, which finds the distribution for the unobserved variables, given the known values for the observed variables and the current estimate of the parameters, and a Maximization...*

Even though CiteSeer enables the visualization of the connection between related articles, it does not provide the user with automatic classification of the *type* of this connection. CiteSeer opted to be non-interpretative, objective, but unhelpful to the user; the user always has to read the citation context in order to work out the relationships.

Nanba and Okumura (1999) introduce a support tool for writing surveys which categorizes citations in text (on the basis of cue words) into "Type C" citations (contrasts), "Type B" citations (based-on relationship) and "Type O" citations (others); Type "C" links are used to display differences and similarities between documents in a *reference graph*. This is a potentially useful way to structure search results, but clusters of papers are often uninformative to users if there is no indication what is similar between papers in this cluster. Users also need to know what single papers are about in "absolute" terms, and not just in relation to other papers—which is typical summary information.

Human-written summaries, on the other hand, do not typically include information about connectedness of research—guidelines actively discourage abstractors from including information about related work. Cremmins (1996) states that it should not be included in an abstract unless the studies are replications or evaluations of earlier work (p. 15). Weil et al. (1963) tell us explicitly never to mention earlier work.

It is our idea that information about connections between papers and local information about one paper should be connected. This could result in a new type of document surrogate which would support the explorative navigation of articles. The processes of search, text skimming and relevance decision could thus be interleaved: during search, parts of a retrieved paper are highlighted; while the reader is navigating the set of returned papers, she might skim-read some of these paragraphs. These text pieces can either directly satisfy the searchers' needs, spark off a new search in a new direction, or convince her that the paper is not relevant after all.

Note how different this relevance decision in such an interactive search-and-display environment is from relevance decision in the paper-based world. There the outcome of the relevance decision was not to be seen for a long time: by the time the paper copy of a certain paper finally arrived, researchers might have half forgotten what their specific reasons for ordering it actually were. Due to this long-term character of relevance decisions, errors were difficult to amend retrospectively, and the risk of ordering the wrong paper was much higher.

Manual summaries are a construct of the paper-based world: texts were of high textual quality, but they were also long-lived and thus fixed. The type of document surrogate we propose will be more dynamic and flexible to the user and her search situation; it should allow for different abstracts to be generated dynamically when needed. Such document surrogates will have a much shorter life span than a valuable human-crafted summary. Even though they will be of lower *textual* quality when compared to such summaries, we predict many situations in which they will have an edge over traditional summaries.

The document surrogate should also include information about similarities and differences between papers; this information could be used either to provide typed links in a citation analysis tool or to enrich the generated summaries.

2.2. Automatic Abstracting

The current state of the art in automatic abstracting is characterized by a deep tension between robustness and depth of understanding. Like machine translation, summarization has been an early target for automation (Luhn, 1958), but the expectation that this is a "easily manageable task" was not fulfilled.

Since the early 90s, with computing power and storage orders of magnitude more plentiful, knowledge-poor, statistical techniques have become fashionable again. However, the view of the complexity of the task has changed within the community. Researchers today see automatic summarization as "one of the most complex tasks of all natural language processing." (Hovy and Lin, 1999, p.92).

Comprehension-based summarization, the traditional symbolic approach, is the most ambitious model for creating automatic summaries. One view is that there cannot be any summarization without a complete comprehension of the text at hand. The argumentation is simple: How should we be able to decide what is important in a text unless we have understood the text?

Figure 2.1 exemplifies the standard model for summarization by comprehension (Spärck Jones, 1994)). It comprises three steps: a) linguistic analysis of the text (syntactic, semantic, pragmatic), which results in the reconstruction of the document semantics in a representation language, b) compression of the contents, by some kind of manipulation of the representation language and finally c) generation of the summary text from the reduced representation.

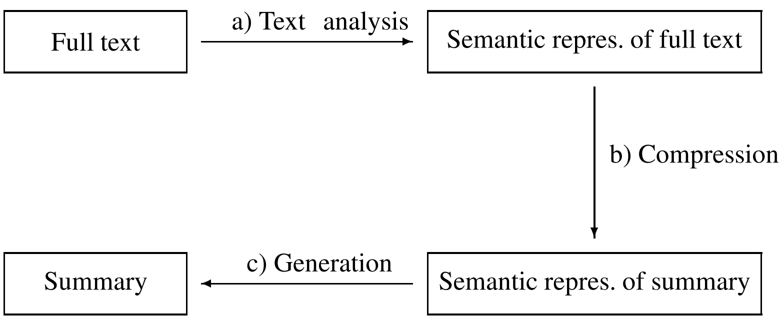

Figure 2.1: Summarization by Text Comprehension

The main problem with this approach is step a): it is not possible yet to map unrestricted text reliably and robustly into a semantic representation. Only then could one apply inference and the other operations that would take place in step b), e.g. following suggestions by Kintsch and van Dijk (1978); Alterman (1985); Brown and Day (1983) and Sherrard (1985). However, severe problems in linguistic analysis and knowledge representation (also referred to as the natural language bottleneck and the artificial intelligence bottleneck) make this model unrealistic for unrestricted text. As a result, people have been looking at alternatives for step a).

Text extraction is one of these alternatives. In this paradigm, step a) is performed in a radical way—each textual segment is condensed to a minimal representation, namely a number of features associated with the textual segment, e.g. whether or not the sentence contains the cue phrase *"to summarize"*. The determination of the features is typically performed in a shallow way, e.g. by calculating the lexical frequency of words in the textual segment, without the use of any linguistic knowledge. Step b), content selection, is performed by selecting a set of these scores, typically the n highest-ranking ones. Step c) is circumvented completely: the outcome of text extraction is the unchanged textual segments whose scores were chosen in step b).

The other solution is based on *fact extraction*. The representation "language" used is a set of frame-like templates (DeJong, 1982; Schank and Abelson, 1977). Step a) is performed by choosing the right template which describes the text, and by filling the slots in the template, e.g. by pattern matching operations. Step b) can be left out completely if the information contained in the templates is already little enough to make up the summary. Otherwise, condensation heuristics decide which ones of several template slots or whole templates are most relevant. Step c), the transformation of the reduced templates into natural language, can be performed either by using fixed templates or by deep generation.

We will in the following look at these two approaches in turn.

2.2.1. Text Extraction

Most of today's summarization systems use text extraction methods, including many commercially available ones, e.g. Microsoft's AutoSummarize (Microsoft, 1997), Oracle (Oracle, 1993), InXight (InXight, 1999) and ProSum (British Telecom, 1998).

The general idea of text extraction is the identification of a small number of "meaningful" sentences or larger text segments from the source text. The most common unit of text extraction is the sentence (Brandow et al., 1995; Kupiec et al., 1995), but some current systems extract paragraphs (Strzalkowski et al., 1999; Abracos and Lopes, 1997; Salton et al., 1994b).

Operational measurements of importance are based on algorithmically determinable properties of the text segment. Each text segment in the source text is scored according to this measure of importance, and subsequently the highest-rated segments are selected.

This produces *extracts* rather than abstracts: collections of the N most "meaningful" text units (sentences), taken verbatim from the text, and presented to the user in the order in which they appeared in the source text.

Extracts can be useful in a document retrieval environment instead of human-written indicative abstracts. A few well-chosen sentences can tell the reader about the terminology used, about the style and syntax, and about how loosely and coherently the text is written. If all the user needs is a tool for rapid relevance assessment, then such robust but uninformed methods can readily provide extracts which meet, to a reasonable degree, the information compression rates required (around 10% of the original text).

Over the years there have been many suggestions as to which low-level features can help determine the importance of a sentence in the context of a source text, such as stochastic measurements for the significance of key words in the sentence (Luhn, 1958; Baxendale, 1958), location of the sentence in the source text (Baxendale, 1958), connections with other sentences (Skorochod'ko, 1972; Salton et al., 1994a), cohesion (Morris and Hirst, 1991; Barzilay and Elhadad, 1999), co-reference information (Baldwin and Morton, 1998), sentence length (Kupiec et al., 1995), the presence of bonus/malus words (Luhn, 1958; Pollock and Zamora, 1975), title words (Edmundson, 1969), proper nouns (Kupiec et al., 1995) or indicator phrases (Paice, 1981; Johnson et al., 1993).

Single heuristics tend to work well on a certain type of document, but in that case success is concentrated on single documents that resemble each other in style and content. For the more robust creation of extracts, e.g., from texts with a high degree of variation in style, it is advantageous to combine these heuristics. The difficulty is to weigh the relative usefulness of single heuristics out of a given set. Edmundson assigns the weights manually. Kupiec et al. (1995) pioneered corpus-driven summarization research in which the combination of heuristics is learned from a training corpus and feature weights are automatically adjusted.

Kupiec et al.'s system uses supervised learning to determine the characteristic properties of those sentences which are known *a priori* to be extract-worthy (positive training examples). The features considered are: presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words (document specific frequency of noun pairs) and occurrence of proper names. They redefine sentence extraction as a statistical classification task: the task is to estimate an unseen sentence's probability to occur in the summary, given its feature values and a statistical model of abstract worthiness acquired during training.

The big advantage of text extraction methods is that they are extremely robust. Due to the low level of analysis performed, it is possible to process texts of all kinds, independent of writing style, text type and subject matter. This means that unexpected turns in a news story, sudden changes in topic and other difficult phenomena can be treated in a shallow way—the extracts will, to a certain degree, reflect these particularities of the texts.

How does one measure the quality of extracts, and what lower bound (baseline) should they be compared to? Researchers have used either random choice of *n* sentences, or selected the *n* leading sentences. Which baseline makes more sense is text type dependent. Brandow et al. (1995) report that for newspaper text, a baseline defined by leading sentences can prove to be so hard to beat that more sophisticated sentence extractors perform *below* the baseline. The reason for this is that journalistic writing style already takes relevance into account by placing the most important information first. For scientific articles, a selection of leading sentences would not make an equally good baseline. Kupiec et al.'s baseline was constructed by leading sentences, and their best results achieved a 74% improvement over baseline. However, with baselines as weak as these, a look at the concrete output is needed to assess the quality of text extracts.

In order to have a concrete example of a sentence extract of a document for ongoing discussion, we used the commercial software AutoSummarize to create extracts of an example article taken from our corpus. This example article—cmp_lg:9408011—will be used throughout the thesis. It is the article most frequently cited by other articles in our collection. The full text of the article is reproduced in appendix B.2 (p. 285). We produced a 10-sentence AutoSummarize extract of the pdf version of the example article, which is given in figure 2.2.

Normally, AutoSummarize displays extracted sentences highlighted in the context where they were extracted from, but it is also possibly to list only the extracted sentences.

AutoSummarize, like many sentence extractors, extracts material other than full document sentences, e.g. titles and headlines (shown in bold face in figure 2.2).

It also selected a single line from the reference list at the end, namely item j), which is the title of a paper published by Rose et al. (1990). This paper is important for the article, but the titles of cited works are no standard summary items, especially

a) **Distributional Clustering of English Sentences**

b) **Distributional Similarity** To cluster nouns n according to their conditional verb distributions pn, we need a measure of similarity between distributions.

c) We will take (1) as our basic clustering model.

d) In particular, the model we use in our experiments has noun clusters with cluster memberships determined by p(njc) and centroid distributions determined by p(vjc).

e) Given any similarity measure d(n;c) between nouns and cluster centroids, the average cluster distortion is

f) If we maximize the cluster membership entropy

g) **Clustering Examples**

h) Figure 1 shows the five words most similar to the each [sic] cluster centroid for the four clusters resulting from the first two cluster splits.

i) **Model Evaluation**

j) 1990. Statistical mechanics and phrase transitions in clustering.

Figure 2.2: AutoSummarize Summary for Example Paper cmp_lg 9408011

if they are not signalled to the user as such. AutoSummarize did not extract sentences from the original abstract, even though the abstract was included in the full document.

In general, extracts are texts of low readability and text quality (Brandow et al., 1995). However this particular AutoSummarize extract reads surprisingly well: it contains no syntactic incoherences like dangling anaphora. None of the selected sentences is obviously displaced in the extract, and they give an idea of the general topic of the paper. We get the idea that it is about clustering, that it is a statistical, technical paper, and that it probably gives an algorithm of some kind. In a document retrieval scenario, this extract could be of use as a rough-and-ready relevance indicator.

Incorrect or confusing content characterization is a harder problem than superficial syntactic flaws, which is why Minel et al. (1997) propose independent evaluation of automatic abstracts by a) text quality and b) content characterization. Even if—like in our extract—each individual sentence is interpretable in isolation, that still does not mean that the extract as a whole will be easy to understand. Earl (1970) noted that extracts are often logically discontinuous. Problems with semantic coherence include unexpected topic shifts or repetitions, non-natural use of anaphora, and general logical incoherence.

With respect to the semantic connection between the sentences, apparent coherence of extracts can even be a *disadvantage*. Sentence d) in the extract appears 25 document sentences after sentence c)—it certainly does not elaborate on particulars related to sentence c). However, as readers are intuitively trying to coerce coherence for prose-like text, they will try to fill in the semantic gaps between potentially unconnected sentences by performing inference (Kintsch and van Dijk, 1978). Many of these inferences might introduce inappropriate semantics links and confuse the reader. In order to avoid this, many summarizers including AutoSummarize offer the possibility to show the extracted sentences highlighted in their original context; others present their extracts as a itemized list with bullet points (Kupiec et al., 1995) instead of continuous prose.

The other issue concerns the extent to which the extract characterizes the meaning of the document. The level of analysis performed seems too low to guarantee correct characterization, and Boguraev and Kennedy (1999) state:

> The cost of avoiding the requirement for a language-aware front end is the complete lack of intelligence—or even context-awareness—at the back end. The validity, and utility, of sentence-or paragraph-sized extracts as representations for the document content is still an open question [. . .]
>
> (Boguraev and Kennedy, 1999, p. 100)

Semantic incoherence and content selection problems become worse the longer the source document is. Typical sentence extractors compress a text down to about 15–25% of the original length—for example, they reduce a short newspaper article to a few sentences. In that case, the extract is still short enough to be read as an indicative "summary", even if the extracted sentences do not form a coherent text. However, things look different for scientific articles, which are much longer. With methods as untargetted as sentence extraction, one needs a 20% compression (or better still, 30%), in order to understand what a text is about: Morris et al.'s (1992) experiment showed that there is no difference in reading comprehension between subjects using the full text, subjects using indicative human-written summaries and subjects using extracts of 20% and 30% compression.

But this level of compression is very low. A 20-page article would have to be reduced to a 4 to 6-page collection of extracted sentences. Given that the statements in such a collection are semantically unconnected, it would be too much text to read and certainly not adequate for human consumption.

One might argue that sentence extracts are a good starting point for later automatic post-processing. However, text extraction is a completely context-insensitive method. Once the abstract-worthy sentences have been extracted, the logical and rhetorical organization of the text is lost. As a result, it becomes difficult to make sensible decisions on how to further reduce a long list of sentences without further information about the meaning of the sentences, the relationships between them or the contexts in which they occurred.

In sum, the low level of analysis performed and its context-insensitivity make text extraction a weak, albeit general and robust technique. Spärck Jones (1999) compares text extraction to looking at a text through tinted glass. All parts of the text can be "seen" by the text summarization technique, but the information we get is certainly blurred.

2.2.2. Fact Extraction

Summarization methods relying on fact extraction need a template to represent the information extracted. We will first discuss the style of these templates and then turn to the question of how to generate coherent summaries from them.

A large-scale competitive evaluation of systems for fact extraction from real-world news paper text was provided by the Message Understanding Conferences (MUC), sponsored by DARPA since the late 1980s (Grishman and Sundheim, 1995). Processing in MUC is restricted to text from a narrow domain, as figure 2.3 shows.

Competition	Domain
MUC 1 & 2	Naval sightings and engagements
MUC 3 & 4	Terrorist attacks in Central and South America
MUC 5	International joint ventures and electronic circuit fabrication
MUC 6	Changes in company management
MUC 7	Telecommunications satellite launches

Figure 2.3: Domains of Texts in Different MUC Competitions

MUC templates are shallow knowledge representation schemes without recursion, which encode information about entities and their relations. They are an instance of the frames well-known from symbolic text understanding and memory organization theories (Minsky, 1975; Schank and Abelson, 1977).

What can summarizers do with such templates? The SUMMONS system as described in Radev and McKeown (1998) and McKeown and Radev (1995) is based

MESSAGE: ID	**TST-REU-0001**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 3, 1996 11:30
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 3, 1996
INCIDENT: LOCATION	Jerusalem
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: 18" "wounded: 10"
PERP: ORGANIZATION ID	

MESSAGE: ID	**TST-REU-0002**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 07:20
PRIMSOURCE: SOURCE	Israel Radio
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 10" "wounded: 30"
PERP: ORGANIZATION ID	

MESSAGE: ID	**TST-REU-0003**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 14:20
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 13" "wounded: more than 100"
PERP: ORGANIZATION ID	"Hamas"

MESSAGE: ID	**TST-REU-0004**
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 4, 1996 14:30
PRIMSOURCE: SOURCE	
INCIDENT: DATE	March 4, 1996
INCIDENT: LOCATION	Tel Aviv
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: at least 12"
"wounded: 105"	
PERP: ORGANIZATION ID	"Hamas"

Figure 2.4: Examples of MUC-4-Style Templates

on deep generation. SUMMONS' speciality is that it compresses several descriptions about the same event from multiple news stories. It takes MUC-4 style templates as input, e.g. the templates given in figure 2.4 (taken from Radev and McKeown 1998, pp. 487-488; the corresponding original newspaper texts are reproduced in figure 2.5). The compression strategy in SUMMONS is specific both to the domain (terrorist activities) and to the text type and situation (journalistic writing, publishing at successive times):

- *Change of perspective:* If the same source reports conflicting information over time, report both pieces of information.
- *Contradiction:* If two or more sources report conflicting information, choose the one that is reported by *independent* sources.

TST-REU-0001
JERUSALEM - A Muslim suicide bomber blew apart 18 people on a Jerusalem bus and wounded 10 in a mirror-image of an attack one week ago. The carnage by Hamas could rob Israel's Prime Minister Shimon Peres of the May 29 election victory he needs to pursue Middle East peacemaking. Peres declared all-out war on Hamas but his tough talk did little to impress stunned residents of Jerusalem who said the election would turn on the issue of personal security.

TST-REU-0002
JERUSALEM - A bomb at a busy Tel Aviv shopping mall killed at least 10 people and wounded 30, Israel radio said quoting police. Army radio said the blast was apparently caused by a suicide bomber. Police said there were many wounded.

TST-REU-0003
A bomb blast ripped through the commercial heart of Tel Aviv Monday, killing at least 13 people and wounding more than 100. Israeli police say an Islamic suicide bomber blew himself up outside a crowded shopping mall. It was the fourth deadly bombing in Israel in nine days. The Islamic fundamentalist group Hamas claimed responsibility for the attacks, which have killed at least 54 people. Hamas is intent on stopping the Middle East peace process. President Clinton joined the voices of international condemnation after the latest attack. He said the "forces of terror shall not triumph" over peacemaking efforts.

TST-REU-0004
TEL AVIV (Reuters) - A Muslim suicide bomber killed at least 12 people and wounded 105, including children, outside a crowded Tel Aviv shopping mall Monday, police said. Sunday, a Hamas suicide bomber killed 18 people on a Jerusalem bus. Hamas has now killed at least 54 people in four attacks in nine days. The windows of stores lining both sides of Dizengoff Street were shattered, the charred skeletons of cars lay in the street, the sidewalks were strewn with blood. The last attack on Dizengoff was in October 1994 when a Hamas suicide bomber killed 22 people on a bus.

Figure 2.5: Articles Corresponding to Templates in Figure 2.4

- *Addition:* If additional information is reported in a *subsequent* article, include the additional information.
- *Refinement:* Prefer more specific information over more general one (name of a terrorist group rather than the fact that it is Palestinian).
- *Agreement:* Agreement between two sources is reported as it will heighten the reader's confidence in the reported fact.
- *Superset/Generalization:* If the same event is reported from different sources

and all of them have incomplete information, report the combination of these pieces of information.

- *Trend:* If two or more messages reflect similar patterns over time, these can be reported in one statement (e.g. three consecutive bombings at the same location).

- *No Information:* Report the lack of information from a certain source when this would be expected.

New templates are generated by combining other templates. The most important template, as determined by heuristics, is chosen for generation.

The content planner assigns values to realization flags (McKeown et al., 1994) related to discourse features such as "similarity" and "contradiction" which guide the choice of connectives and control local choices such as tense and voice in later generation steps. These switches also govern the presence or lack of certain constituents, in order to satisfy anaphora constraints and to avoid repetition of constituents. SUMMONS uses a domain ontology for lexical choice, to enrich the input and to make generalizations. The sentence generator used is FUF (Elhadad, 1993; Robin, 1994) which employs SURGE, a large systemic grammar of English. The output of this process is the following summary:

> *Reuters reported that 18 people were killed in a Jerusalem bombing Sunday. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel Radio. Reuters reported that the radical Muslim group Hamas had claimed responsibility for the act.*

The fact that this summary is deep-generated is illustrated by the change of voice in the first sentence compared to its source (TXT-REU-0001), the change of tense in the third sentence from simple past to past perfect, the replacement of the phrase *"the Islamic fundamentalist group Hamas"* by *"the radical Muslim group Hamas"* (TXT-REU-0003) and the occurrence of the term "*the next day*" which did not appear in the original text, but was added by SUMMONS during the combination and surface realization phase.

A similar, but more surface-oriented approach is given in Paice and Jones (1993) for scientific papers in the field of crop husbandry. The slots in their template (cf. figure 2.6, taken from Paice and Jones 1993, p. 71) are also domain specific,

	Paper 1	Paper 2
SPECIES:	*potato*	*winter wheat*
CULTIVAR:		
HIGH LEVEL PROPERTY:	*yield*	*each field a grid*
LOW LEVEL PROPERTY:		
PEST:	Powdery mildew	*Brent Geese Branta*
AGENT:		
INFLUENCE:		
LOCATION:	*York, Lincoln and Peterbourgh, England*	*Deepsdale Marsh, Burnham, Deepdale*
TIME:		*1985, 1986*
SOIL:		
CLIMATE:		
TREATMENT:		
PROCESS:		
NUTRIENT:		

Figure 2.6: Paice and Jones' (1993) Template for Agricultural Articles

Paper 1:

Title: The assesment [sic] of the tolerance of partially resistant potato clones to damage by the potato cyst nematode Globodera pallida at different sites and in different years.

Ann. Appl. Biol., 1988, 113:79-88

This paper studies the effect the pest G. pallida has on the yield of potato. An experiment in 1985 and 1986 at York, Lincoln and Peterbourgh, England was undertaken. These results indicate clearly *that* there are consistent differences between potato cultivars in their tolerance of damage by PCN as measured by proportional yield loss.

Paper 2:

Title: The effect on winter wheat of grazing by Brent Geese *Branta Bernicla*

Journal of Applied Ecology, 1990, 27:821-833

This paper studies the effect of Brent Geese Branta on the each field a grid of winter wheat [sic]. The experiment took place at Deepdale Marsh, Burnham, Deepdale. The fact that ear density increased due to grazing in one yield indicates that there is probably little value in the farmer sowing seed at a higher density in an attempt to compensate for geese grazing.

Figure 2.7: Paice and Jones' (1993) Abstracts for the Papers in Figure 2.6

e.g. SPECIES, CULTIVAR and PEST. The concepts are identified by a heuristic pattern matching procedure, where patterns such as "*effect of* INFLUENCE *on* PROPERTY *of/in* SPECIES" are identified in text. Candidate strings for a certain slot are weighted according to their frequency and the contexts where they appeared. Oakes and Paice

(1999) introduce an automated process to generate the search patterns automatically from text.

The abstracts, cf. figure 2.7 (taken from Paice and Jones 1993, p. 74), are generated in a much simpler fashion than Radev and McKeown's. The first sentence in each abstract is generated by slotting the best candidate strings into a fixed natural language template. Note that when a wrong string has been identified, such as the string *"each field a grid of"* in the second abstract, this might lead to ungrammatical output. The second sentence in each abstract is added by traditional text extraction: if a phrase like "*results indicate that*" (underlined in figure 2.7) is encountered, the sentence is added, in the hope that this turns the abstract into an informative one.

In fact-extraction templates, domain-knowledge is hard-wired into the slot definitions, and semantic relations between the slots are known *a priori*, e.g., the knowledge that it is the PERPETRATOR of a terrorist act who causes the killing or wounding of the HUMAN TARGETS. The depth of representation and the additional knowledge about semantic relationships between slots has clear advantages: it is possible, on the basis of domain-specific templates, to generate high-quality abstracts which read well and which are logically well-structured, as exemplified by Radev and McKeown's and Paice and Jones' summaries.

One of the disadvantages of such domain-specific approaches is the huge knowledge engineering efforts required to hard-wire the knowledge into the recognizers. Worse still, the whole machinery (template filling and, as a result, summarization) is not robust enough to react to unforeseen events in the texts. Only text segments that fit the expectations expressed by the situation slots can be handled. For instance, in the SUMMONS example only those aspects which have been anticipated in the template can be treated in the summary, namely the effects of the attack in terms of physical damage. All the other information in the original text is ignored, e.g. information about Mr. Peres and his prospects in the election (an important part of Text TST-REU-0001), or the future of the peace process and the international reaction to the attack (additional information in Text TST-REU-0003). Paice and Jones can similarly only process articles from a narrow subject field.

Spärck Jones (1999) calls fact extraction methods "what you know is what you get" techniques (p. 2), as they come with "the disadvantages that the required type of information has to be explicitly (and often effortfully) specified and may not be important for the source itself" (p. 3).

In sum, we have seen that the state of the art in automatic summarization is far from creating fluent summaries of unrestricted text which characterize the text's meaning well. However, there are two practical approaches which manage to fulfill some of the requirement of this task. We will in the following suggest our own approach.

2.3. A New Approach

In our review of current abstracting techniques, we found the following requirements for a new type of automatically generated document surrogate:

- It should be more flexible towards the text than fact extract based summaries are, while retaining some of the expressiveness of these.
- It should contain more information than text extracts, while retaining some of the generality and robustness of these.
- It should be more adaptive with respect to other tasks and other users than manual summaries are, while retaining the good characterization of the article achieved by these.
- It should include types of information not typically occurring in manual summaries (e.g. related work and its relation to the current work), while integrating this information with all other aspects.

2.3.1. Design of the Approach

2.3.1.1. General Design Criteria

When designing a new document surrogate, we started from the requirement of robustness. Robustness is indeed imperative, as we are working with unrestricted, naturally occurring text; such "real-life" text is a rough species. As a direct result, we decided to take orthographic sentences as unit of annotation, in analogy to most text extraction methods. Sentences can be identified robustly; smaller units seem fraught with problems. The concept of a clause, for example, has had linguists arguing for a long time.

Of course, a document surrogate based on textually extracted sentences presupposes that sentences which can act as parts of summaries are indeed found in the document, as Radev and McKeown (1998) point out. If this is not the case, nothing but deep-generation will help. However, we assume that explicit material for summaries will be available, due to the authors' motivation to formulate their important claims clearly.

One of our central observations is that the importance of a sentence within the whole text is crucially influenced by its rhetorical status: depending on whether the sentence describes the purpose of the research, the conclusion, or the author's criticism of other research, the content of a given sentence might be more or less useful for a given information need. For example, sentences which describe weaknesses of previous research can provide a good characterization of the scientific articles in which they occur, since they are likely to also be a description of the problem the paper is intending to solve. Take a sentence like *"Unfortunately, this work does not solve problem X"*: if X is a shortcoming in somebody else's work, the sentence might be a very good candidate for extraction. However, a very similar-looking sentence can play a completely different rhetorical role: if X refers to limitations of the approach presented in the paper, the sentence is *not* a good characterization of the article at all.

Our novel contribution is that we attach *additional rhetorical information* to the extracted sentences, in the form of fixed labels. The purpose of the labels is to capture the global context in which the sentence occurred in respect to the overall argumentation in the document. In contrast to fact extraction methods, the semantics of these labels is not defined by domain-specific knowledge, as this was the reason for the inflexibility which plagues fact extraction methods. This is in the line of Kircz (1991) and Sillince (1992) who have argued that *rhetorical* (or argumentative) indexing will provide more domain-independence in document retrieval applications than semantic indexing does. The exact definition of the labels will be given in section 2.3.2 and justified in chapter 3. As a result of how the labels are defined, they should apply equally well to articles coming from different disciplines; the approach is thus *domain independent but text type dependent*.

Some of these labels we define will encode different types of *connections* between articles: contrastive vs continuative mentions of other work, as motivated in section 2.1.2. The advantages of such a typing of links become apparent for large volume search, where a pre-sorting by type of link will save the user valuable time. However, the typing is subjective in nature (cf. section 3.2.2). Humans might disagree about certain cases, and a system performing the differentiation will sometimes make errors. We are aware of this risk, but think that the advantages outweigh the risks. Additionally, we invest some effort to measure the subjectivity of such decisions.

It is the working hypothesis of this thesis that shallow *argumentative* analysis is a promising approach for document characterization in a document retrieval environment. We take the deliberate decision not to model the *scientific content* of the article—in contrast to other approaches, which shallowly model content by term frequency methods (Salton et al., 1994b), lexical chaining methods (Baldwin et al., 1998; Barzilay and Elhadad, 1999), TextTiling (Hearst, 1997) or lexical similarity (Kozima, 1993). One of the reasons for our decision is the observation that even in human summarization it is not always the case that knowledge-intensive methods are the method of choice. Cremmins (1996) states that professional abstractors do not attempt to fully "understand" the text, but use surface-level features such as headings, key phrases and position in paragraphs. They also use *discourse* features such as overall text structure to organize abstracts and extract information. Endres-Niggemeyer et al. (1995) found that they

- prefer top-level segments of documents,
- build topic sentences,
- consider beginnings and ends of units as relevant,
- examine passages and paragraphs before individual sentences,
- exploit document outlines,
- pay attention to document formatting,
- determine the role of each section in the problem-solving process by reading the first and last sentence of each section or each paragraph and
- paraphrase relations between theme and in-text summaries.

However, our emphasis on the rhetorical side of the analysis does not mean that we believe that domain knowledge should never be included in a summarizer for scientific articles. On the contrary, scientific knowledge about the contents of the articles is undoubtedly going to improve the overall summarization process. Our long-term vision is that a better system would incorporate both *form* and *content* approaches, as we expect them to complement each other perfectly by recovering different aspects of meaning in the article. However, given the state of the art, we feel it is currently most promising to use shallow approaches of *form* rather than *content*.

The fundamental question, of course, is the question of depth of analysis, to which we will return in detail in chapter 5. Our approach will opt for robust, low-level techniques, because we believe that many of the problems encountered can be successfully addressed with fairly shallow techniques. Our approach is corpus-based: we will observe or learn features from a large amount of naturally occurring text. In sum, our approach

- uses shallow analysis;
- relies on sentences as units of extraction and analysis;
- does not model scientific content;
- attaches rhetorical information to sentences, e.g. the type of relation to other work.

The document surrogate we sketched so far bears comparison to structured abstracts, as sentences are classified into different types of information. Therefore, we will now review the literature on structured abstracts.

2.3.1.2. Structured abstracts

The literature on abstracting has identified the following four *content units* for informative summaries of articles in the experimental sciences (ANSI, 1979; ISO, 1976; Rowley, 1982; Cremmins, 1996):

- PURPOSE/PROBLEM
- SCOPE/METHODOLOGY
- RESULTS
- CONCLUSIONS/RECOMMENDATIONS

There is more disagreement about “peripheral” content units, such as RELATED WORK, BACKGROUND, INCIDENTAL FINDINGS and FUTURE WORK. According to Alley (1996), BACKGROUND is a useful content unit in an abstract if it is restricted to being the first sentence of the abstract (p. 22). Other authors (Rowley, 1982; Cremmins, 1996) recommend not to include any background information at all. Similar disagreement concerns the content unit RELATED WORK, as already discussed.

Buxton and Meadows (1978) provide a comparative survey of the contents units in summaries in the physics domain. They studied which rhetorical section in the source text (*Introduction–Method–Result–Discussion*) corresponds to the information in the summaries and found, for example, that summaries tend not to report material from the *Method* section. Milas-Bracovic (1987) performed a similar experiment on sociological and humanities summaries. Tibbo (1992) compares science (chemistry), social science (psychology) and humanities (history) with respect to the following content categories: BACKGROUND, PURPOSE/SCOPE, HYPOTHESES, METHODOLOGY, RESULTS, and CONCLUSIONS. Although the ANSI standard claims applicability of the above-mentioned four information units for abstracting in the social sciences and humanities as well, she found that fewer than 40% of the sentences in the history summaries fell into one of the ANSI categories.

Some innovative approaches suggest completely new information units and new structures. Trawinski (1989) introduces *problem structured abstracts*, with the main categories DOCUMENT PROBLEM, PROBLEM SOLUTION and TESTING METHOD, RELATED PROBLEMS, and 63 more fine-grained content elements such as SPECIFICATION OF OBJECTS USED IN TESTING and POSSIBLE USAGE AREAS IN SCIENCE. Broer (1971) uses graphic block-like units in his two-dimensional summaries, with the following units: WHAT? TITLE, WHAT/WHY? – INSTRUMENT, WHAT/WHY? – PRELIMS, WHAT? – CONSTRUCTION, HOW? – BASIC, HOW? – AID and WHY? – PERFORMANCE. His approach sounds promising but has not been used in practice.

Liddy (1991) showed experimentally that professional abstractors use an internalized building-plan when they write summaries. Her description of the components of summaries of empirical articles is based on professional abstractors' intuitions and a corpus of summaries.

Figure 2.8 gives an overview of the components (taken from Liddy 1991, p. 71). The seven most important components ("*prototypical components*") are displayed in capitals and bold face. The next level of importance ("*typical components*") is shown in capitals. The components found by Liddy cover short text spans (parts of sentences rather than sentences) and they can be embedded recursively into each other. Liddy concludes that abstractors, even if they might not choose the same sentences, still choose the same *type* of contents when they fill the fixed building-plans.

In the medical field, structured abstracts (Adhoc, 1987; Rennie and Glass, 1991) have long replaced free text summaries. Abstract information is given using prescribed headings which are dependent on the type of research being reported. Rather elaborate rules for their preparation have been established (cf. for example,

- background
  - RELATION TO OTHER RESEARCH
  - new terms defined
  - institution
  - administrators
  - location of study
- **PURPOSE**
  - **HYPOTHESIS**
    - independent variable
    - dependent variable
  - research questions
  - RESEARCH TOPIC
  - **SUBJECTS**
    - SAMPLE SELECTION
    - control population
  - no of experiments
  - time frame
- **METHODOLOGY**
  - PROCEDURES
    - CONDITIONS
    - materials
  - DATA COLLECTION
  - data analysis
- **RESULTS**
  - reliability
  - DISCUSSION
    - unique features
    - limitations
- **CONCLUSIONS**
  - significance of results
  - IMPLICATIONS
  - practical applications
  - future research needs
- appendices
  - **REFERENCES**
  - tables

Figure 2.8: Liddy's (1991) Empirical Summary Components

Haynes (1990)). The following headings are used for descriptions of clinical trial reports in the *Annals of Internal Medicine*: BACKGROUND, OBJECTIVE, DESIGN, SETTING, PATIENTS, INTERVENTIONS, MEASUREMENTS, RESULTS and CONCLUSIONS. For reviews, headings include OBJECTIVE DATA SOURCES and STUDY SELECTION. Summaries in the *Archives of Dermatology* (Arndt, 1992) are structured into: BACKGROUND/DESIGN, RESULTS, CONCLUSIONS (CLINICAL), BACKGROUND/OBSERVATIONS and CONCLUSIONS (OBSERVATIONAL).

Several researchers found problems with the application of structured abstracts. Salager-Meyer (1992) researches empirically the linguistic and discoursal quality of

Task	Information required
Browsing the Literature	OBJECTIVES and CONCLUSIONS of a clinical study
Evaluating Clinical Studies	EXPERIMENTAL DESIGN plus CONCLUSIONS of the research (STUDY TYPE, STATISTICS, LIMITATIONS)
Matching Patients with Clinical Studies	ELIGIBILITY AND EXCLUSION CRITERIA, EXPERIMENTAL SETTING
Treating/Counseling Patients	INTERVENTIONS, RISK FACTORS, DIAGNOSTIC TESTS, ADVERSE EFFECTS and CONCLUSIONS
Planning Clinical Research	OBJECTIVE, CONCLUSIONS, DISCUSSION of UNANSWERED QUESTIONS and FUTURE WORK, LIST OF REFERENCES

Figure 2.9: ACP's Annals Extracts: Tasks and Components

medical summaries, in connection to content units. She found almost half to be "poorly structured", i.e. discoursally flawed. Froom and Froom (1993) showed that structured abstracts in *Annals of Internal Medicine* do not always contain all of the information requested in the guidelines for authors, even when the information needed was present in the article itself.

However, Hartley et al. (1996) and Hartley and Sydes (1997) present experiments which give evidence that structured abstracts are easier to read and overall more efficient than prose summaries. Hartley (1997) argues that structured abstracts should also be applied to social sciences. Taddio et al. (1994), based on a larger study of 300 summaries from three journals, also found that the structured abstracts were more likely to contain more complete information of research importance than unstructured abstracts were.

A new summarization/extraction application in the medical domain tests the plausible assumption that task flexibility can be realized based on such content units: the American College of Physicians (ACP) has recently started providing *task-specific* summaries for the papers in *Annals of Internal Medicine* (ACP online, 1997; Wellons and Purcell, 1999). There is a choice of five different types of (manually created) extracts for each paper; each of the five types is geared towards a different medical tasks. These tasks have been identified as frequently recurring in the different types of professional work of the readership of the *Annals*. Each of these tasks requires a different type of information from the medical articles, cf. figure 2.9.

And finally, Buckingham Shum and colleagues propose a specific meta data scheme for expressing relationships between articles (Shum, 1998; Sumner and Shum, 1998; Shum et al., 1999). It is a meta-data scheme for a Scientific Knowledge Web (SKW) of scientific papers in the field of HCI (Human–Computer Interaction) which concentrates on scholarly discourse, and the expression of relations between papers. The status of the units of this document surrogate is not anchored in any scientific domain knowledge, but rather in higher-level aspects which connect the instances of research, e.g. similarities and differences between scientific approaches. We will take the same approach in the design of our document surrogate. Their suggestion is unusual in its emphasis on *relations* between pieces of research, another aspect which has inspired the design of our document surrogate. An example for a representation of a paper according to this meta-description can be seen in figure 2.10 (taken from Shum 1998, p. 19).

There are 10 relations which describe how scientific works might be related to each other: ANALYSES, SOLVES, DESCRIBES-NEW, USES/APPLIES, MODIFIES/EXTENDS, CHARACTERIZES /RECASTS, EVALUATES (SUPPORTS or PROBLEMATISES or CHALLENGES).

The suggested concepts are entities which are important in the domain (HCI), namely the following 9 categories: APPLIED-PROBLEM, THEORETICAL-PROBLEM, METHOD, LANGUAGE, SOFTWARE, EVIDENCE, THEORY/FRAMEWORK, TREND,

REF: Smith, J. (1997) ATC Overload, Journal of ATC, 3 (4), 100-150		
ANALYSES	APPLIED-PROBLEM	*Air traffic controller cognitive overload*
USES/APPLIES	THEORY/FRAMEWORK	*use of video, undergraduate university physics, student ability*
PROBLEMATISES	SOFTWARE	*GOMS cognitive modelling tools*
MODIFIES/EXTENDS	LANGUAGE	*Knowledge Interchange Format (KIF)*
CHARACTERIZES/RECASTS	TREND	*Electronic trading over the internet*
CHALLENGES	SCHOOL-OF-THOUGHT	*Postmodernism*
SUPPORTS	EVIDENCE	*multimedia, school chemistry teaching*

Figure 2.10: Shum's (1998) Design for Document Representations in a Scientific Knowledge Web (SKW)

SCHOOL-OF-THOUGHT. Each of the concepts can be further refined by keywords or names and connected to a reference or a URL.

The design of the SKW slots has not been verified by cognitive experiments with users, but is currently in a beta-testing phase, where researchers in the HCI field can contribute example encodings of their own papers, suggestions and comments. In the setup that Shum (1998) has in mind, a human expert would select one of these possible slots and fill them manually with domain-specific material, sometimes requiring background knowledge and inference. This is typical for meta-data approaches, which assume in general that humans (authors or indexers) provide mark-up. Shum (1998) argues pessimistically about the task of filling the slots in his scheme by an automatic process:

> It is possible that useful information may be extracted through intelligent analyses of text, but often this information is not explicit in documents, but implicit in the minds of domain experts. (Shum, 1998, p. 16)

On the one hand, we welcome the meta-data approach because meta-indexing provided by authors can be expected to be of high quality. On the other hand, it might take some time before such meta-data approaches will have an impact on writer's behaviour when papers are written and submitted.

The main difference between our design and this scheme is the fact that our analysis is aiming to provide filling material *automatically*. As a result, the fillers which our planned document representation provides have to be of a much simpler kind: mere surface strings.

Another difference is that in Shum's approach nodes themselves are "neutral" (i.e., not associated with local semantic information); the only semantics that a node has comes from the links and its position in a research web. In our approach, the characterization of the paper on its own is also important. This has the advantage that papers can be summarized and characterized as single items without looking at their connections (which the system does not necessarily have knowledge of).

2.3.2. Rhetorical Document Profiles (RDPs)

The outcome of these design decisions is a new document surrogate. We call this document surrogate a *Rhetorical Document Profile* (RDP) because it consists of rhetorical units (slots) and because it profiles different kinds of information about the document. RDPs were designed to encode typical information needs of new readers in a systematic and structured way. Figure 2.11 shows an empty RDP.

1. Solution Identifier	—	
2. Specific Aim/Scope	—	
3. Background	Aim	Problem/Phenomenon
	—	—
4. Solution/Inventive Step	—	
5. Claim/Conclusion	—	

Rel. to Other Work	6. Rival/Contrast	Reference	Solution ID	Type of Contrast
		[...]	—	—
		[...]	—	—
	7. Basis/Continuation	Reference	Solution ID	Type of Continuation
		[...]	—	—
		[...]	—	—

External Structure	Headlines	8. Textual Structure
	—	—
	—	—

Figure 2.11: An Empty Rhetorical Document Profile (RDP)

On the following pages, we will walk the reader through a *filled* RDP (namely the one for example article cmp_lg/9408011) slot by slot. This RDP was manually filled by us with textual material taken verbatim from the source article (excluding the human-written summary). These surface strings are often whole sentences, and sometimes segments of sentences. Slot fillers are identified by sentence numbers, which act as pointers into the original text where the textual material was extracted from (cf. sentence numbers in XML representation of the article, appendix B.1).

The exact filling criteria will be elaborated later. The solution displayed is *one* possible solution; as the filling criteria rely on human intuition, other solutions would have been possible too. We claim, however, that other humans would have filled the slots sufficiently *similarly*; chapter 4 will provide experimental evidence for this claim.

1. SOLUTION IDENTIFIER	—

SOLUTION IDENTIFIER: Sometimes a paper introduces a new approach and gives it a name. Later papers might refer to it using that term. In our domain, these are often artefacts: names of programs, methods, algorithms or theories. Information about well-known methods in the field is extremely important to uninformed and partially informed readers (cf. section 1.1). Examples for what we will consider as identifiers for solutions are the following: "*the SPLATTER parser", "Maximum Entropy classifier", "Minimum Description Length (MDL)", "Data Oriented Parsing (DOP)", "the Centering algorithm*" and *"Rhetorical Structure Theory (RST)*". A solution identifier does not always have to be a proper name, but can be any other description, e.g. "*Hobbs' anaphora resolution algorithm*" or "*simulated annealing*".

Our example article does introduce a named solution: a new method which later articles refer to as "*soft word clustering*". But unfortunately, there is no explicit mention of this particular term in the example article itself. A similar expression ("*hierarchical "soft" clustering*") does appear in the author-written summary, but we decided not to use information from the summary. As it is, the slot remains empty.

2. SPECIFIC AIM/SCOPE

164	to group words according to their participation in particular grammatical relations with other words
10	how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves
44	how to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams
11	how to derive the classes directly from distributional data
46	learning a joint distribution of pairs from a large sample of pairs
22	we will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs
45	we will only address the more specific problem in which the objects are nouns and the contexts are verbs that take the nouns as direct objects

The slot SPECIFIC AIM/SCOPE contains descriptions of the research goal specific to the article. We believe that fillers of this slot can be the single most characteristic information about a scientific paper (particularly if they occur in a sentence together with the methodology used).

Our example article happens to contain unusually many explicit mentions of the specific research goal. The slot-fillers differ in the level of abstraction at which they describe the research goal, and in their focus on a particular aspect of the problem. Some of them are paraphrases of each other, or contribute more detailed information. This leads to a certain degree of redundancy. Note that slot fillers **11** and **46** do not just talk about the research goals, but additionally give some information about the solution, i.e., *how* the task is solved. In general, it can be difficult to keep goals and solutions apart. Slot fillers **22** and **45** stand in the context of a contrastive *scope* delimitation: the authors stress that they do *not* classify verbs, just nouns.

3. BACKGROUND	
AIM	PROBLEM/PHENOMENON
1 automatically classifying words	**4** The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

BACKGROUND information divides into two kinds: BACKGROUND (AIM) can be considered as the paper's topic, a high level characterization of the task, e.g. "*machine translation*". In our example, the high level goal is the automatic classification of words. BACKGROUND (PROBLEM/PHENOMENON) gives high level problems in the field (in this case: data sparseness). If the paper aims at an explanatory account, then BACKGROUND (PROBLEM/PHENOMENON) can contain sentences describing phenomena to be explained.

---

4. SOLUTION/INVENTIVE STEP

164 a general divisive clustering procedure for probability distributions can be used [. . .]
12 we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.

The nature of the SOLUTION/INVENTIVE STEP depends on the type of discipline we are considering. In some empirical disciplines, a new empirical claim or a new hypothesis is the main innovation of the paper; the research goal, namely to verify or disprove the hypothesis, is left implicit. In those disciplines, the methodology is often standardized. In disciplines like computational linguistics, the main idea is often the technical solution (methodology) — exactly because there are few fixed rules as to which methodologies can be used.

In our case, there are some high-level descriptions of the innovative step: the authors apply a well-known general divisive clustering procedure, and part of their solution is to model word senses as clusters.

---

5. CLAIM/CONCLUSION

165 The resulting clusters are intuitively informative, and can be used to construct class-based word coocurrence models with substantial predictive power.

The CLAIM/CONCLUSION slot concerns explicit claims. Explicit claims, hypotheses and predictions are typically found in experimental papers. Even though this particular paper is a technical paper (something is engineered), we still encounter a claim. This claim, however, is not a claim about the scientific domain, but rather a meta-claim: it is a statement that the problem has been solved, and that the result makes sense. Such sentences, if correctly identified, can give valuable information about the paper's problem-solving process.

Two slots describe the relation of the current work to other work. The two categories are 6. CONTRASTIVE relations and 7. CONTINUATION of research relations. The slot RIVAL/CONTRAST approaches is filled with information on other work which is in a contrastive or comparative relationship to the given work, or information about a specific weakness of the other work. The other work can be identified either by a formal explicit reference or by a solution *identifier*, in analogy to the SOLUTION IDENTIFIER slot discussed on p. 58.

6. RIVAL/CONTRAST

REFERENCE	SOLUTION ID	TYPE OF CONTRAST
• [Hindle 1990] – **5**		**9** it is not clear how it can be used directly to construct word classes and corresponding models of association
• [Brown et al. 1992] – **13**	**13** other class-based modeling techniques	**13** Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information
• [Resnik 1992] – **11**		**11** preexisting sense classes (Resnik) vs. we derive the classes directly from distributional data
•	**43** agglomerative clustering techniques	**43** need to compare individual objects being considered for grouping (advantage of authors' method)
• [Church and Gale 1991] – **40**	**40** smoothing zero frequencies appropriately	**41** However, this is not very satisfactory as our goal is to avoid the problems of data sparseness by clustering words together

With respect to contrastive approaches, the authors seem to have identified certain weaknesses with Hindle's (1991) and Brown et al.'s (1993) work. There is also a contrast in task with Resnik (1992), and an advantage over both agglomerative clustering techniques and Church and Gale's (1991) approach.

7. BASIS/CONTINUATION

	REFERENCE	SOLUTION ID	TYPE OF CONTINUATION
•	[Rose et al. 1990] – **113**	**113** deterministic annealing	**113** The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering [Rose et al. 1990] ...
•	[Dagan et al. 1993] – **155**		**155** based on a suggestion by
•		**29** Kullback-Leibler (KL) distance	**29** used
•	[Hindle 1993] – **19**		**19** automatically parsed by Hindle's parser
•	[Church 1988] – **20**		**20** with the help of a statistical part-of-speech tagger
•	[Yarowsky 1992] – **20**		**20** [with the help of] tools for regular expression pattern matching on tagged corpora

The BASIS/CONTINUATION describes work which provides a starting point for the current work, or which provides data, theoretical apparatus or methodology that the current work uses. It might also support the claims of the given paper, or fit in with the paper's claims without contradiction. Information about intellectual ancestry, i.e., the knowledge of who builds their work on who else's work, is of great importance to users trying to orient themselves in a new area (cf. section 1.1). Note that contrasted and continued research are not necessarily mutually exclusive classes. Researchers might use a certain work as starting point but identify problems with it which they then try to rectify.

In the example paper, the single most important continuation is the fact that the authors use Rose et al.'s annealing procedure. They also use Hindle's (1993) parser, Church's (1988) POS tagger, Yarowsky's (1992) regular expression tools and a commonly agreed upon statistical measure (KL). Also, they use a suggestion in a paper by Dagan et al. (1993).

EXTERNAL STRUCTURE

HEADLINES		8. TEXTUAL STRUCTURE
1.	Introduction	
1.1	Problem Setting	
1.2	Distributional Similarity	
2.	Theoretical Basis	
2.1	Distributional Clustering	
2.1.1.	Maximum Likelihood Cluster Centroids	
2.1.2.	Maximum Entropy Cluster Membership	
2.1.3.	Minimizing the Average KL Distortion	
2.1.4.	The Free Energy Function	
2.2.	Hierarchical Clustering	
3.	Clustering Examples	**127** All our experiments involve the asymmetric model described in the previous section.
4.	Model Evaluation	
4.1.	Relative Entropy	
4.2.	Decision Task	
5.	Conclusions	

EXTERNAL STRUCTURE is a concerned with explicit representations of structure in the article: a simple listing of all headlines found in the text (sub-slot HEADLINES) or explicit textual information about the section structure (sub-slot TEXTUAL STRUCTURE). In this paper, only one explicit statement about textual structure was found (and even this one is not a clear case). It is a reference back to the previous section, and can give some indication of the contents of that section.

The full RDP is given in appendix B.3; appendix B.4 lists the sentences from the original text corresponding to the textual material in the RDP.

We have by now redefined the goal of the thesis: to verify if it is possible to automatically identify these types of information in real world texts. The output of this thesis, namely relevant textual material for the RDP slots, could be regarded as a final result. We believe that lists of RDP slot fillers are already better textual extracts than those provided by today's sentence extraction methods. Additionally, we predict that RDP slot fillers would provide useful information for human abstractors, shortening the time it takes them to construct a full textual abstract. Conceptually however, the extraction step described in this thesis was designed in such a way that its output would be of greatest possibly usability to the follow-on processing steps.

We will now discuss the use of RDP type information in a document retrieval environment.

2.3.3. RDPs for Tailored Summaries

If RDPs could be automatically compiled in an off-line fashion for each document in a large collection of papers, this would have definite advantages for document retrieval. RDPs in themselves provide a detailed, tabularized summary of the article. Users could get an overview of the contents of the paper by directly scanning them. However, RDPs are big document surrogates containing a lot of redundancy. Users might not want to invest the time to directly read them.

Users who prefer more traditional summaries could be provided with user, length and task tailored summaries generated from RDPs. Imagine two kinds of users (informed vs. uninformed readers), three kinds of "tasks" (general purpose, contrastive use of summaries, determining intellectual ancestry between papers) and two lengths of summaries (longer vs. shorter). In figure 2.12, simple recipes (or building-plans) for summaries are given for combinations of expertise, length and task. The building-plans vary in the number and type of individual slot fillers which are included in the summary. Following from our considerations in section 2.1.1, the building-plan mirror the following intuitions about differences in expertise:

- More background material (e.g. in the introduction) is needed for uninformed readers, whereas informed readers do not require any background information. For uninformed readers, the approaches of other researchers are *described*; for informed readers, they are only *identified* (by direct citation or by solution identifier).
- This should make summaries for uninformed reader in general longer than summaries for informed readers.
- Sentences with more general terms are preferred for uninformed readers, and sentences with more technical terms for informed readers. Sentence **44** in figure 2.13, which contains for example the specific term *"ngram", "linguistic objects"*, was chosen as expression of the SPECIFIC AIM for informed readers, whereas sentence **164** in figure 2.17 was chosen for uninformed readers, as it contains more general terms (*"group", "words", "grammatical relations"*).

The second factor we considered was *task-tailoring*:

- General purpose summaries consist of as few SPECIFIC AIM sentences as possible, in order to avoid redundancy.

- Longer general purpose summaries should include some SOLUTION/INVENTIVE STEP material, in order to simulate informative summaries.
- For comparative or contrastive summaries, the "most important" rival approaches should be presented to the reader. One simple way to determine importance of an approach is by measuring how much space the description of the approach is given in the paper (see also a later discussion of this point in section 3.4).
- In analogy, the most important based-upon other work needs to be identified for intellectual-ancestry summaries.

We manually generated summaries to illustrate the building-plans. Many ways

	Informed reader	Uninformed reader
General purpose, short	*Summary 1:* 2 SPECIFIC AIM	*Summary 5:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM
General purpose, longer	*Summary 2:* 2–3 SPECIFIC AIM + 1 INVENTIVE STEP	*Summary 6:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2–3 SPECIFIC AIM + 1 INVENTIVE STEP
Contrastive	*Summary 3:* 2 SPECIFIC AIM + 1–2 (SOLUTION ID + TYPE OF CONTRAST)	*Summary 7:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM + 1–2 (DESCR. OF OTHER WORK + TYPE OF CONTRAST)
Ancestry	*Summary 4:* 2 SPECIFIC AIM + 1–2 (SOLUTION ID + TYPE OF CONTINUATION)	*Summary 8:* 1 BACKGROUND (AIM) + 1 BACKGROUND (PROBLEM) + 2 SPECIFIC AIM + 1–2 (DESCR. OF OTHER WORK + TYPE OF CONTINUATION)

Figure 2.12: Building-Plans for Task and Expertise Tailored Summaries

of arriving at the actual summary text are imaginable for this illustration, resulting in summaries of a different quality. We decided to select good candidates amongst the RDP slot fillers and to change them as little as possible. The output is enriched with templates, and some minimal surface repair is performed in order to make the result easier to read.

We simulated a selection process amongst RDP slot fillers for each slot given in the building-plan. The rules for choosing a given sentence for a slot over its competitors are that it has to be a) minimally similar to any other chosen sentence for that slot, in order to reduce redundancy and b) maximally similar to as many other candidates for that slot as possible—which are, as a consequence of a), *not* chosen. The argumentation for this is due to Edmundson (1969) who voiced the intuition that more important material appears redundantly in text. The occurrence of similar slot fillers thus raises our confidence that the given slot fillers are good characterizations for the semantics of its slot.

Surface repair can be imagined as follows: for a summary sentence about research goal, strings are taken from the corresponding RDP slot, the semantic verb is identified and transformed into the syntactic form fitting to the template context (*"This paper's goal is to"*). Template material is shown underlined in the following summaries.

As there is more space for the discussion of other approaches in summaries for uninformed readers, it is not always necessary to process the sentences further. In contrast, generating concise sentences for informed readers is a more complex task, as the material needs to be found from different sources and assembled correctly. Consider, for example, the sentence constructed from sentences **5** and **9** in figure 2.15, where sentence **5** supplies the solution identifier and sentence **9** supplies the criticism/contrast. In order to correctly handle comparison and negation in sentences **5/9** and **14**, some more complex templates or deeper generation mechanisms would have to be used here.

44 <u>*This paper's goal is to*</u> *organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.*
22 <u>*More specifically: the goal is to*</u> *classify nouns according to their distribution as direct objects of verbs.*

Figure 2.13: Summary 1: Informed Reader, General Purpose, Short

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **11** *The goal is to derive the classes directly from distributional data.* **164** *A general decisive clustering procedure for probability distributions is used.*

Figure 2.14: Summary 2: Informed Reader, General Purpose, Longer

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **5** *Unlike, [Hindle 1990],* **9** *this approach constructs word classes and corresponding models of association directly.* **14** *In comparison to [Brown et al. 92], the method is combinatorially less demanding and does not depend on frequency counts for joint events involving particular words, a potentially unreliable source of information.*

Figure 2.15: Summary 3: Informed Reader, Contrastive

44 *This paper's goal is to organize a set of linguistic objects such as words according to the contexts in which they occur, for instance grammatical constructions or n-grams.* **22** *More specifically: the goal is to classify nouns according to their distribution as direct objects of verbs.* **113** *It uses the deterministic annealing procedure introduced by [Rose et al 1990].*

Figure 2.16: Summary 4: Informed Reader, Intellectual Ancestry

1 *<u>This paper's topic is to</u> automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *<u>This paper's specific goal is to</u> group words according to their participation in particular grammatical relations with other words,* **22** *<u>more specifically to</u> classify nouns according to their distribution as direct objects of verbs.*

Figure 2.17: Summary 5: Uninformed Reader, General Purpose, Short

1 *<u>This paper's topic is to</u> automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *<u>This paper's specific goal is to</u> group words according to their participation in particular grammatical relations with other words,* **22** *<u>more specifically</u> to classify nouns according to their distribution as direct objects of verbs.* **11** *<u>Another goal is to</u> derive the classes directly from distributional data.* **12** *<u>The authors</u> model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN> for each word w.*

Figure 2.18: Summary 6: Uninformed Reader, General Purpose, Longer

1 *<u>This paper's topic is to</u> automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *<u>This paper's specific goal is to</u> group words according to their participation in particular grammatical relations with other words,* **22** *<u>more specifically</u> to classify nouns according to their distribution as direct objects of verbs.*
5 *[Hindle 1990] proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen.* **8** *In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events.* **9** *It is not clear how his notion of similarity can be used directly to construct word classes and corresponding models of association.*
13 *Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes [Brown et al. 1990].* **14** *Class construction is then combinatorially very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information.*

Figure 2.19: Summary 7: Uninformed Reader, Contrastive

1 *<u>This paper's topic is to</u> automatically classify words.* **4** *The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.* **164** *<u>This paper's specific goal is to</u> group words according to their participation in particular grammatical relations with other words,* **22** *<u>more specifically</u> to classify nouns according to their distribution as direct objects of verbs.*
113 *<u>The authors use</u> a deterministic annealing procedure for clustering [Rose et al. 1990], in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN/> following an annealing schedule.*

Figure 2.20: Summary 8: Uninformed Reader, Intellectual Ancestry

The summaries read fluently and convey different kinds of information for different readers and different tasks. Manipulation of length and of syntactic constructions in the sentences is possible due to the rhetorical information coming from the RDP slots. This information is not domain-specific, in contrast to similar fact-extraction templates.

Multi-document summarization could also profit from RDPs for scientific articles: articles mentioning similar concepts in the same RDP slots might be candidates for collective characterization in one summary for all these articles. Documents returned by a users' query for the term "*Decision Tree Learning*" might be described ("summarized") as follows:

> *In your query results, there are 13 papers that have the term* `PP attachment` *in their* SPECIFIC AIM *slot. There are 33 papers with* `cross-validation` *in the* SOLUTION *slot.*

2.3.4. RDPs for Citation Maps

The information contained in RDPs can help users understand the relationship of one particular paper to other papers: either to papers contained in a set of search results, or to papers already known to the user.

We suggest generating a new construct called *local citation maps* on the fly for papers of interest. Figure 2.21 shows such a (manually created) citation map, including all those papers from our document collection which cite our example paper, Pereira et al. (1993). Each article of this starting set is displayed in a rectangle and identified by name of authors and year of publication. The map also shows articles referenced by these papers (i.e. those not contained in our document collection) which are displayed without rectangles. (The difference in status between articles within and outwith our collection is of course that we cannot trace the citations contained in the latter.)

The information contained in RDPs allows to display *typed* links, where the green links corresponds to CONTRAST ("contrasting the work to other work") and purple links to BASIS/CONTINUATION ("building the work onto previous solutions"). If no particular stance could be determined, a "neutral" citation link is displayed in black.

We claim that citation maps could help users picture document similarities and differences in an immediate and natural way. Especially for uninformed searchers, such a representation of links would be extremely useful for a local exploration of a wide range of questions.

Certain kinds of similarities and differences between papers can be seen at first glance. Figure 2.21 shows that Nitta and Niwa (1994) and Resnik (1995) cite Pereira et al. (1993) and the other four papers in our collection only contrastively, and they both cite some other papers, and in a contrastive way (e.g. Schütze (1993) and Hirst (1991)). Two of the other three papers, on the other hand, also form a natural sub-cluster: Dagan

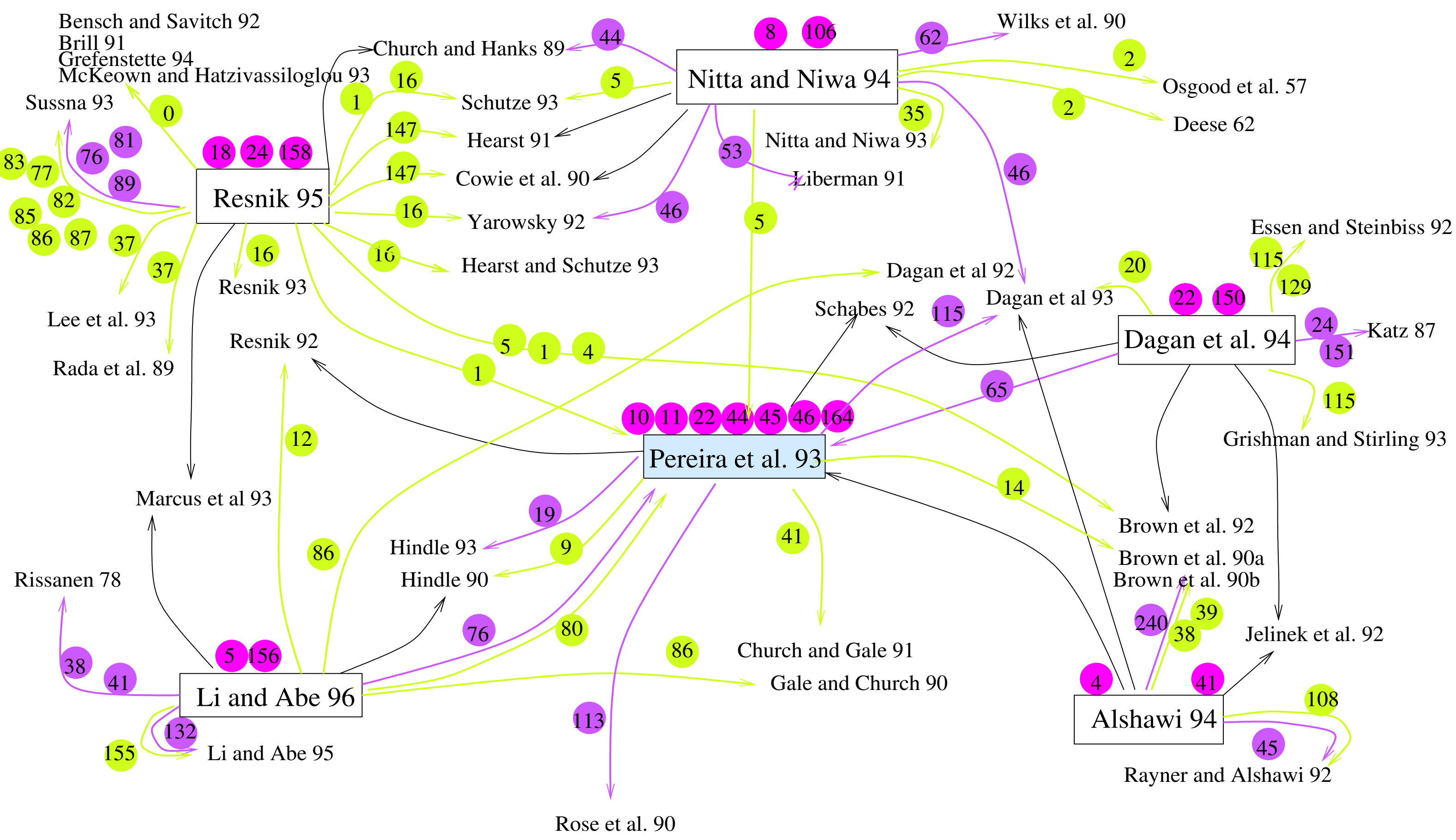

Figure 2.21: Citation Map for Document 9408011

et al. (1994) and Alshawi (1994) cite Pereira et al. (1993) positively or neutrally. Li and Abe (1996) cite Pereira et al. (1993) in both continuation as well as contrast context and have no direct citation relations to any of the other papers.

Citation maps do not give temporal information a privileged status, but information about the time of publication can also be relevant to searches: for example, rival approaches are typically those working in the same time fragment.

More information could be displayed in the citation map by expansion: links could be expanded into full sentences interactively, namely the sentences in the paper which explicitly express a continuation relationship or a contrast (represented by their numbers and coloured circles corresponding in figure 2.21). For example, figure 2.22 shows in which respect Nitta/Niwa, Resnik and Li/Abe contrast themselves to Pereira et al. (1993).

Contrasting paper	Contrast/Criticism
[Nitta and Niwa, 1994]	*However, using the co-occurrence statistics requires a huge corpus that covers even most rare words.* (S-5, 9503025)
[Resnik, 1995]	*However, for many tasks, one is interested in relationships among word senses, not words.* (S-1, 9511006)
[Li and Abe, 1996]	*Here, we restrict our attention on 'hard clustering' (i.e., each word must belong to exactly one class), in part because we are interested in comparing the thesauri constructed by our method with existing hand-made thesauri.* (S-80, 9605014)

Figure 2.22: Contrasting and Criticizing Citations to 9408011 in Other Articles

Whereas Nitta and Niwa's contrasting statement could be seen as a criticism, the other papers point out differences in their *aim or scope*: senses vs. words, or hard vs. soft clustering.

Note the similarity between citation maps and what Bazerman (1985) calls *research maps*: he argues that experienced researchers in a field have organized their knowledge in the field in a kind of linked representation centered around research goals, methodologies, researcher names, research groups and schools (cf. section 2.1.2). A tool that creates citation maps from RDPs would support uninformed users in acquiring their own mental research map more efficiently. Local and content-enriched citation maps present information in an immediate, powerful and natural way.

Uninformed users could start using citation maps without any knowledge of the terminology in the field. They get an overview of relations amongst papers and incidentally come across relevant terms in sentences which are displayed. This boot-strap knowledge will make subsequent keyword searches more efficient.

2.4. Conclusion

In this chapter we have looked at state-of-the-art summarization techniques. An overview of the paper-based world of hand-written summaries has shown that such summaries are of high quality but inflexible. They also do not provide much-needed information about contrastive and ancestral relations between similar articles. With respect to automatic summarization, we found that fact extraction methods, while providing informative output, are too domain-dependent and not robust enough towards unexpected turns in unrestricted texts—whereas text extraction methods, which are robust to the extreme, do not provide enough information about the extracted material. We have argued that what is missing is some form of context with respect to the overall document content. As a possible way out of this predicament, this chapter has introduced RDPs (Rhetorical Document Profiles).

- Similar to *text-extraction* methods, RDPs will use sentences as extraction units. In contrast to text-extraction output, RDPs contain information attached to each sentence, namely the information about the rhetorical status of a sentence with respect to the whole paper. This makes different kinds of postprocessing possible.

- Similar to *fact-extraction* approaches, summaries can be (re)generated, due to the information connected with the textual material. In contrast to fact-extraction templates, RDP slot semantics are not domain dependent: RDP slots do not encode anything about the subject matter of science. However, RDP slots are text type dependent.

- Similar to *human-written abstracts*, information about functional units in the document will help construct and structure the abstract in an RDP-based approach. In contrast to human-written summaries, RDPs provide information about connections between articles; they can be tailored to user expertise and task requirements.

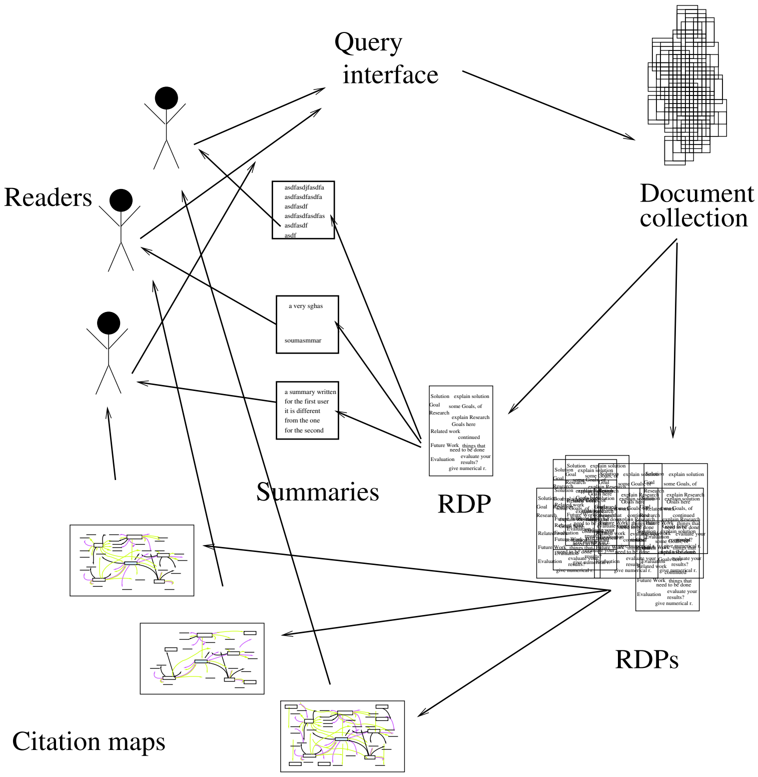

Figure 2.23: The Role of RDPs in a Document Retrieval Environment

- Similar to *citation-indexing tools*, RDPs provide information about relatedness of articles. In contrast to them, RDPs distinguish the *type* of links between documents and also provide *static, semantic* information about the document.

As figure 2.23 shows, RDPs could support scientists' information foraging activities in an actual document retrieval environment by providing the information needed for automatically generated, expertise and task tailored summaries and for citation maps.

This thesis will not go all the way in producing RDPs automatically—RDPs are highly informative document surrogates, the automatic generation of which is too ambitious a task for the scope of this thesis. Instead, this thesis will constitute the first step in the production of RDPs, namely the production of a list of sentences which are good slot fillers for RDPs.

In this context, the next chapter will place the concept of an RDP (which is a reader-centered construct) with the concept of argumentative zones in text (which is a writer-centered construct). It will pave the way for an automatic procedure for filling RDP slots, by looking at strategies for finding good slot fillers in running text.

Chapter 3

Argumentative Zoning

In the previous chapter, we motivated a new document surrogate, the RDP or rhetorical document profile. We showed that the RDP is a desirable construct in a document retrieval environment, as it provides the right kind of information for the flexible generation of summaries.

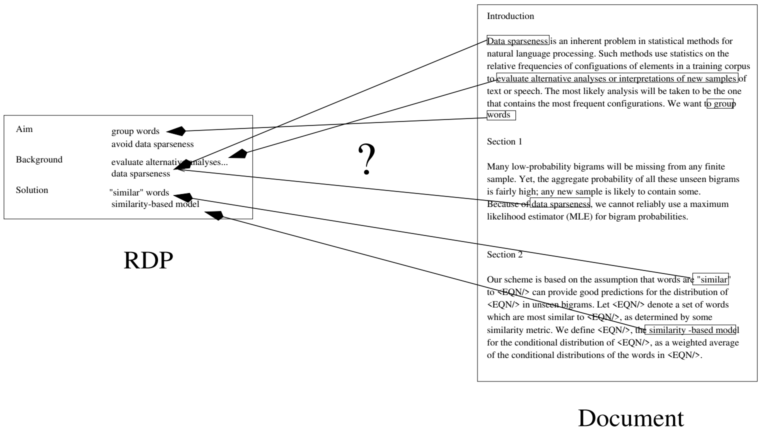

Figure 3.1: From Documents to RDPs

In this chapter we discuss how to get from text to RDPs. Some constraints of the task were already discussed at the end of the previous chapter: our analysis will be shallow and robust, using full sentences as filling material, and it will aim at attaching rhetorical information to the extracted sentences (cf. figure 3.1).

In the previous chapter, the semantics of RDP slots was justified by the document retrieval task: the slots are defined by the kinds of information that *readers* want *out* of the text. In this chapter, we will define the slot semantics by looking at what the *writer* put *into* the text, in particular how she organized and structured her text. This has a parallel to the situation in summarization in general, about which Paris writes:

> Summarising depends on the recognition of both the intention of the writer in writing the original text (with respect to what he or she was trying to convey) as well as the goals and knowledge of the reader (why do they want a summary and how much do they know about the domain). (Paris, 1993, p. 1)

However, it is not obvious *what* kind of rhetorical information should define the slot semantics. We will see in section 3.1 that fixed section structure cannot offer much help. We base our structural analysis instead on a new model of prototypical scientific argumentation. The theory behind the model, described in section 3.2, is based on authors' communicative acts—-these communicative acts are predictable from text type-specific expectations. The model draws from different strands of research:

- *Argumentative moves:* Swales (1990) claims that there is a restricted range of prototypical argumentative goals that a writer of a scientific article has to fulfill, e.g., to convince her readers that the problem she addresses has some interest to the field (cf. section 3.2.1).

- *Authors' stance towards other work:* The field of Content Citation Analysis categorizes semantic relations between citing and cited work (cf. section 3.2.2).

- *Intellectual ownership:* Authorship in scientific discourse is typically explicitly given: either the statements are presented as own work, as well-known facts in the field, or as other authors' claims. We will argue in section 3.2.3 that a segmentation based on this distinction is an essential step for our task. To our knowledge, this aspect of scientific text has not received any attention in computational approaches yet.

- *Problem-solving statements:* Scientific research papers can be seen as biased reports of a problem-solving activity: they contain many statements about problem-solving activities: own as well as other researchers' (cf. section 3.2.4). Some of these problem-solving activities are portrayed as successful, others as flawed.

Our model of scientific argumentation is operationalized in section 3.3, where we introduce our practical annotation scheme and the task of *Argumentative Zoning*, i.e. the task of applying the scheme to text. Section 3.4 makes the connection back to RDPs and shows how Argumentative Zoning serves the construction of RDPs.

The task introduced in this chapter, Argumentative Zoning, is new, but fits in with the recent surge of interest in document profiling, argumentation and discourse analysis. We will contrast Argumentative Zoning with related work in section 3.5.

3.1. Fixed Section Structure

RDP slots are in many cases identical with the common section headings in scientific articles. The task of filling the slots would be simplified a great deal if we knew from which section in the paper to extract the corresponding material.

The single most prominent property which is the same across many scientific articles is their common external global structure in *rhetorical sections* (or *rhetorical divisions*) and corresponding section headers (van Dijk, 1980). This highly structured building plan for research articles is particularly well-established in the life and experimental sciences, e.g. experimental physics, biology and psychology. The most famous structure is four-pronged and contains the sections *Introduction, Method, Results, Discussion*. In some disciplines, there is a fifth typical section, namely *Conclusions*. Rhetorical sections often contain other rhetorical sections, e.g., a *Method* section in a psychology article is often divided into *Subjects, Materials* and *Procedure*. Rigid section structures enhance efficiency of understanding and information searching: researchers in psycholinguistics, for example, know with great accuracy where to find the number of experimental subjects in any given article.

It has been argued that this structure has evolved and become petrified because texts which serve a common purpose among a community of users eventually take on a predictable structure of presentation (Mullins et al., 1988; Hyland, 1998).

Knowing how to write in this style is important for the career of scientists, but they are rarely trained in it during their undergraduate degrees. Part of the training of young researchers consists in experienced researchers showing them "how to write papers such that they get accepted". Rules on how to fit material into sections do exist (e.g., "*report only numerical results in the* RESULTS *section; if there's interpretation involved, put it into the* DISCUSSION *section", "description of machinery belongs into the methodology except if...* "). Prescriptive style manuals and writer aids abound (Mathes and Stevenson, 1976; Blicq, 1983; Alley, 1996; Conway, 1987; Day, 1995;

Farr, 1985; Houp and Pearsall, 1988; Michaelson, 1980; Mitchell, 1968; van Emden and Easteal, 1996; Lannon, 1993). Writing style manuals urge writers to explicitly mark explicit structure, e.g.:

- by clear physical format/layout: orthographically recognizable indications of text structure;
- by mapping of conceptual paragraphs to physical paragraphs;
- by use of informative sub-headings as very short summaries;
- by adherence to conventionalised text structure;
- by explicit signalling of text macrostructure (*"in section 2, we will ..."*);
- by clear discourse/rhetorical relations;
- by clear and logical elaboration of the subject matter (topicality and nuclearity).

There have been more or less formal attempts by discourse analysts to model this section structure. Van Dijk (1980) presented conventionalized schematic forms for several text types (apart from experimental research reports, also for narratives, arguments, newspaper articles).

Figure 3.2 shows Kircz' (1991) taxonomy of argumentative entities (taken from Kircz 1991, p. 368), which is more fine-grained than van Dijk's, and specifically designed for physics articles. It also includes dependencies between these entities in the form of see-also links and in the form of logical implications (i.e., there cannot be any experimental constraints if there is no experimental setup), which we have not reproduced here. This structure, though it covers the whole article, is similar to Liddy's structured abstract and other abstract templates. Kando (1997) presents a similar structure which she uses to make queries in a DR environment more distinctive, cf. figure 3.3, taken from (Kando, 1997, p. 70).

Models such as Kando's and Kircz' describe papers from the experimental sciences well. However, our corpus covers an interdisciplinary science. In cognitive science and computational linguistics, where the focus is the investigation and simulation of intelligent action and language processing, a wide range of scientific areas is covered: experimental sciences (psychology, neuroscience), engineering (computer

1. Definition of the research subject in broad terms
   (a) Redefinition of the problem in the actual research context
2. Experimental setup
   (a) Experimental constraints
   (b) Experimental assumptions
   (c) Experimental ambiguities
   (d) Relation of experimental setup with other experiments
3. Data collection
   (a) Data handling methods
   (b) Data handling criteria
   (c) Error analysis
4. Presentation of raw experimental data
   (a) Presentation of smoothed experimental data
   (b) Pointers to pictorial or tabular presentation
   (c) Comparison of own data with other results
5. Theoretical model
   (a) Theoretical constraints
   (b) Theoretical assumptions
   (c) Theoretical ambiguities
   (d) Relation of theoretical elaboration with other works
6. Theoretical/mathematical elaboration
7. Presentation of theoretical results/predictions
   (a) Comparison with other theoretical results
   (b) Pointers to pictorial or tabular presentation
8. Comparison of experimental results with own theoretical results
   (a) Comparison of experimental results with other theoretical results
   (b) Pointers to pictorial or tabular presentation
9. Conclusions
   (a) Experimental conclusions
   (b) Theoretical conclusions
10. Reference to own previous published work
    (a) Reference to own work in progress
11. Reference to other people's published work
    (a) Reference to other people's work in progress

Figure 3.2: Kircz' (1991) Argumentative Taxonomy

- A. PROBLEMS
  - A.1 BACKGROUND
    - A.1.1 stating background WITHOUT REFERENCES
    - A.1.2 REVIEW or relevant previous research
  - A.2 RATIONALE
    - A.2.1 GAP of knowledge
    - A.2.2 IMPORTANCE
    - A.2.3 INDUCEMENTS to start the study
    - A.2.4 INTERESTS OF THE AUTHOR(S)
  - A.3 RESEARCH TOPIC
    - A.3.1 RESEARCH QUESTIONS
      - A.3.1.1 HYPOTHESIS
      - A.3.1.2 PURPOSES
    - A.3.2 SCOPE of the study
      - A.3.2.1 OUTLINE of methods
      - A.3.2.2 OUTLINE of discussion
      - A.3.2.3 principle RESULT or conclusion
      - A.3.3.4 ORGANIZATION of the paper
  - A.4 TERM DEFINITION
- B. VALIDITY of the evidence or METHODS
  - B.1. FRAMEWORK of the study
    - B.1.1. RESEARCH DESIGN
    - B.1.2. ENVIRONMENT
    - B.1.3. MODELS/ASSUMPTIONS used in the study
    - B.1.4. REASONS for selecting the framework
  - B.2. SUBJECTS
    - B.2.1. ATTRIBUTES of the subjects
    - B.2.2. SELECTION CRITERIA of the subjects
    - B.2.3. NUMBERS of the subjects
    - B.2.4. REASONS for selecting the subjects
    - B.2.5. ETHICAL CONTROLS for the subjects
  - B.3. OPERATIONS/inventions
    - B.3.1. PROCEDURES of the operation
    - B.3.2. TOOLS used in the operation
    - B.3.3. MATERIALS used in the operation
    - B.3.4. CONDITIONS of the operation
    - B.3.5. REASONS for selecting the operation
  - B.4. DATA COLLECTION
    - B.4.1 PROCEDURES and ITEMS of the data collection
    - B.4.2. TOOLS used in the data collection
    - B.4.3. MATERIALS used in the data collection
    - B.4.4. CONDITIONS of the data collection
    - B.4.5. MEASUREMENT CRITERIA
    - B.4.6 REASONS for selecting the data collection
  - B.5. DATA ANALYSIS
    - B.5.1 PROCEDURES and TECHNIQUES of analysis
    - B.5.2. TOOLS and S/W used in the data analysis
    - B.5.3. REASONS for selecting the analysis
  - B.6. LOGICAL EXPANSION
- C. EXAMINATION of the EVIDENCE
  - C.1. PRESENTATION OF EVIDENCE
  - C.2. ORIGINAL EVIDENCE, mentioned again
  - C.3. ORIGINAL EVIDENCE + opinion
  - C.4. ORIGINAL EVIDENCE + SECONDARY EVIDENCE
  - C.5. ORIGINAL EVIDENCE + SECONDARY EVIDENCE + OPINION
  - C.6. SECONDARY EVIDENCE
  - C.7. SECONDARY EVIDENCE + OPINION
  - C.8. OPINION
- E. ANSWERS
  - E.1. SUMMARY of the study
  - E.2. CONCLUSIONS
  - E.3. FUTURE RESEARCH
  - E.4. APPLICATIONS
  - E.5. SIGNIFICANCE

Figure 3.3: Kando's (1997) Categories

science, language engineering, artificial intelligence), humanities (philosophy), sociology (sociology of science), applied sciences (discourse analysis, English for a Specific Purpose), medicine and theoretical sciences (linguistics and mathematics). The scientific traditions of the authors represented in our corpus vary according to many dimensions:

- *Structure:* In contrast to experimental scientists, humanists comply much less to the classic model for scientific writing (Tibbo, 1992). Tibbo states that the contents of humanistic writing frequently appear as seemingly unstructured text lacking standardized section headings. Historical discourse, for example, consists mainly of interpretative arguments and narrative supporting those arguments.

- *Research style:* In young disciplines new methods evolve fast, as researchers use and combine relatively new techniques with old and new tasks. Additionally, new disciplines often have not agreed on what a good evaluation strategy is. An example for this is the current state of the field of automatic summarization.

- *Cultural differences:* Different language traditions prefer different argumentative structure, as has been shown in the case of English–German (Clyne, 1987) and Polish–English (Duszak, 1994). The main difference seems to be that in the German-Polish tradition the results are kept "hidden" as long as possible, in order to retain the readers' curiosity, whereas the English texts preview the structure of the entire article and give results away early.

- *Conference and Presentation style:* The presentation of a paper can be influenced by how conferences are organized. In philosophy, speakers read their talks from paper, whereas in linguistics free talks prevail, supported by handouts. In computational linguistics, computer science and psychology, where talks are also free, there are printed proceedings and no handouts. In neuroscience, however, talks are often accompanied by a slide show.

- *Peer reviewing:* Researchers in interdisciplinary fields often have to review papers with material coming from a discipline adjacent to their own. They typically do not feel that they should criticize the presentation of that material. As a result, there is a general leniency towards writing style; papers with diverging structure *are* accepted at conferences and in journals.

As predicted, the structure of the papers in our corpus is indeed heterogeneous. Even though most of our articles have introduction and conclusions sections (sometimes occurring under headers with different names), the presentation of the problem and the methodology/solution are idiosyncratic to the domain and personal writing style. In some cases, prototypical headers are used, in others, headers contain subject-matter terms.

Figure 3.4 compares the most frequent headlines in our corpus (left hand side) with those in a comparison corpus of cardiology papers. 74% of all 823 headers in our data are not prototypical. 32% of all papers contain no explicitly marked *Conclusion* section. In the entire CL corpus, there were only two sections titled *Method* or *Methods*.

Computational Linguistics (80 papers)			Cardiology (103 papers)		
Headline	Frequency		Headline	Frequency	
Introduction	63	79%	Introduction	103	100%
Conclusion	34	43%	Results	97	94%
Discussion	13	16%	Discussion	97	94%
Conclusions	13	16%	Methods	95	92%
Acknowledgments	12	15%	Tables	81	79%
Results	8	10%	Statistics	41	40%
Experimental Results	8	10%	Patients	30	29%
Evaluation	7	9%	Limitations	29	28%
Background	7	9%	Conclusions	26	25%
Implementation	6	8%	Statistical Analysis	23	22%
Example	6	8%	Conclusion	18	17%
Acknowledgements	6	8%	Patient Characteristics	9	9%

Figure 3.4: Frequencies of Headlines in CL and Cardiology Corpus

In contrast to the computational linguistics corpus, where the external structure of the paper if obviously a matter of personal style, the section structure in the medical corpus is very homogeneous: each headline out of the typical *Introduction, Method, Result, Discussion* structure is present in almost each paper. The least frequent component, *Methods*, is still present in 92% of all papers. Some papers (25%) contain a *Conclusion* section as a fifth section structure. The only headings that were not prototypical occurred at a deeper level of embedding (e.g. names of specific medical procedures or methodologies such as "*Measurement of lipid hydroperoxides*").

Of course, rhetorical sections in our data might still be present logically even if they are not explicitly marked. In the absence of an *Introduction* section, the same function is sometimes fulfilled by sections titled *Motivation* or *Background*, or by the first paragraphs of the first section. However, in this case it is much harder to find the corresponding types of information.

Overall, if section structure is not the dominant structure in our data, we will have to consider other possible commonalities between the papers. The variation in our data forces us to steer clear of distinctions that are too domain specific. We will have to go “deeper” into the structure of the papers—we believe that more interesting theoretical questions will emerge this way. However, due to the robustness requirements of our approach, we cannot go indefinitely deep: the commonalities we are looking for must still be traceable on the surface.

3.2. A Model of Prototypical Scientific Argumentation

3.2.1. Argumentative Moves: Swales (1990)

We have so far presented scientific articles as purpose-free, objective descriptions of research. The rigid section structure reinforces the impression that the research presented was performed following a strictly logical procedure. However, the process by which a scientific paper is created is very complex—there are many levels of actions that interact, presentational as well as scientific (Latour and Woolgar, 1986). The presentation of research in scientific papers does not normally follow the chronological course of the research. Ziman (1969) states that the authors do not inform of false starts, mistakes, unnecessary complications, difficulties and hesitations. On the contrary, the procedure is shown as simple, precise, profitable and the conclusions derived as inevitable. If we accept a definition of argument as “any proof, demonstration, or reason that is useful for persuading the audience of the validity of a statement” (Myers, 1992), then arguing is an important part of presenting science, even in disciplines where overt argumentation is not part of the presentational tradition.

Swales (1990) assumes that the main communicative goal authors of scientific papers is to convince readers of the validity and importance of their work, as this is the only way to have the paper reviewed positively, and published as a result. Authors need to show that the presented research is justified (i.e., that it addresses an interesting problem), that it is a contribution to science, that the solution presented is a good solution, and that the evaluation is sound.

His CARS model ("Creating a Research Space") describes the structure of introductions to scientific articles according to prototypical rhetorical building plans. The unit of analysis is the argumentative *move* ("a semantic unit related to the writer's purpose"), typically one clause or sentence long. There is a finite number of such moves, and they are subdivided into "steps". The model, a successor of his earlier model (Swales, 1981), is schematically depicted in figure 3.5. It is based on empirical studies on two data collections: firstly, a collection of several hundred research articles in the physical sciences and secondly, a mixed collection of research articles from several science and engineering fields.

One such rhetorical move is to motivate the need for the research presented (Move 2), which can be done in different ways, e.g. by pointing out a weakness of a previous approach (Move 2A/B) or by explicitly stating the research question (Move 2C). Note that context plays an important role for the classification of a sentence in Swales' model: the example sentence for Move 2C (which characterizes the question actually addressed in the article) would constitute a different move if it had appeared towards the end of the article, e.g. under the heading *Future Work*.

Swales' model has been used extensively by discourse analysts and researchers in the field of English for Specific Purposes, and for tasks as varied as teaching English as a foreign language, human translation and citation analysis (Myers, 1992; Thompson and Yiyun, 1991; Duszak, 1994). Salager-Meyer (1990, 1991, 1992) establishes similar moves for medical abstracts. Busch-Lauer (1995) did not find these moves in all abstracts of her German medical corpus; she concludes that presentation and arrangement of moves are related to the author's intentions and summarizing skills.

An inspection of introduction sections in our corpus showed that Swales' definition of argumentative moves seem to generalize well to the domain of computational linguistics and cognitive science. (Crookes (1986), however, reports that is not the case for the social science literature.) As a result of the shortness of our texts, however, the optional move 3.3 (INDICATE ARTICLE STRUCTURE) was rare. The right hand side of figure 3.5 shows real examples coming from our corpus.

Even though Swales' model is non-computational, i.e. not aimed at automatic recognition of the moves, one important assumption in Swales' work is that the argumentative status of a certain move is visible on the surface by linguistic cues. This is important for our task.

We will use a description based on argumentative moves to describe structural similarities between papers in our corpus, but we feel that we cannot use Swales' model

MOVE 1: ESTABLISHING A TERRITORY		
1.1	CLAIMING CENTRALITY	• *Recently, there has been a lot of interest in Earley deduction* [...] (S-0, 9502004)
1.2	MAKING TOPIC GENERALIZATIONS (BACKGROUND KNOWLEDGE) OR	• *The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept.* (S-3, 9503002)
	(DESCRIPTION OF PHENOMENA)	• *In the Japanese language, the causative and the change of voice are realized by agglutinations of those auxiliary verbs at the tail of current verbs.* (S-56, 9411021)
1.3	REVIEWING PREVIOUS RESEARCH	• *Brown et al. (1992) suggest a class-based n-gram model in which words with similar cooccurrence distributions are clustered in word classes.* (S-12, 9405001)
MOVE 2: ESTABLISHING A NICHE		
2A	COUNTER-CLAIMING	• *I argue that Hidden Markov Models are unsuited to the task* [...] (S-9, 9410022)
or 2B	INDICATING A GAP	• [...] *and to my knowledge, no previous work has proposed any principles for when to include optional information* [...] (S-9, 9503018)
or 2C	QUESTION-RAISING	• *How do children combine the information they perceive from different sources?* (S-15, 9412005)
or 2D	CONTINUING A TRADITION	• *Within a current project on adapting bilingual dictionaries* [...] *the need arose for a POS-disambiguator to facilitate a context sensitive dictionary look-up system.* (S-4, 9502038)
MOVE 3: OCCUPYING A NICHE		
3.1A	OUTLINING PURPOSE	• *The aim of this paper is to examine the role that training plays in the tagging process* [...] (S-32, 9410012)
or 3.1B	ANNOUNCING PRESENT RESEARCH	• *In this paper, we discuss the interaction of temporal anaphora and quantification over eventualities.* (S-2, 9502023)
3.2	ANNOUNCING PRINCIPLE FINDINGS	• *In our corpus study, we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts* [...] (S-139, 9504006)
3.3	INDICATING ARTICLE STRUCTURE	• *This paper is organized as follows: We first review a general algorithm for least-errors recognition* [...] (S-27, 9502024)

Figure 3.5: Swales' (1990) CARS Model; Examples from our Corpus

without adjustment. Firstly, whereas Swales' scheme covers only the introduction we need a model that describes the whole article; some moves might have to be added. Also, many of Swales' definitions are vague. For example, the difference between the two moves 2D (CONTINUING A TRADITION) and 2C (INDICATING A GAP) is that for move 2D "there is a weaker challenge to the previous research" (Swales, 1990, p. 156). Our feeling is that the scheme would need to be operationalized before it could be applied by groups of annotators.

Swales's (1990) model is more flexible than models of fixed section structure like van Dijk's. However, it still assumes an argumentative structure which is rather close to the textual form, with a fixed order of moves. We empirically found that the order he suggests is typically indeed the most frequent, but we also found many cases in our heterogeneous corpus where the argumentative moves were ordered in unexpected ways. For example, six of our texts started with a specific goal statement, and 14 introductions do not contain any explicit goal statement at all. Duszak (1994) reports similar problems with Swales' assumption of a fixed move order.

Swales' move name		Our move name
1.1	Claiming Centrality	DESCRIBE: GENERAL GOAL SHOW: OWN GOAL/PROBLEM IS IMPORTANT/INTERESTING SHOW: SOLUTION TO OWN PROBLEM IS DESIRABLE SHOW: OWN GOAL/PROBLEM IS HARD
1.2	Making Topic Generalizations	DESCRIBE: GENERAL PROBLEM DESCRIBE: GENERAL CONCLUSION/CLAIM
1.3	Reviewing Previous Research	DESCRIBE: OTHER CONCLUSION/CLAIM
3.1A	Outlining Purpose	DESCRIBE: OWN GOAL/PROBLEM
3.1B	Announcing Present Research	DESCRIBE: OWN GOAL/PROBLEM
3.2	Announcing Principle Findings	DESCRIBE: OWN CONCLUSION/CLAIM
3.3	Indicating Article Structure	DESCRIBE: ARTICLE STRUCTURE PREVIEW: SECTION CONTENTS SUMMARIZE: SECTION CONTENTS

Figure 3.6: Move Names in Swales' and in our Model

We borrow Swales' moves given in figure 3.6 and expand them to the moves in figure 3.7. These 12 moves are a useful description of a large part of the material occurring in the introduction sections and some other material too.

The moves for textual presentation (Swales' "Indicate Article Struc-

1. DESCRIBE: GENERAL GOAL
   Abstract generation is, like Machine Translation, one of the ultimate goal [sic] of Natural Language Processing. (S-0, 9411023)
2. SHOW: OWN GOAL/PROBLEM IS IMPORTANT/INTERESTING
   Both principle-based parsing and probabilistic methods for the analysis of natural language have become popular in the last decade. (S-0, 9408004)
3. SHOW: SOLUTION TO OWN PROBLEM IS DESIRABLE
   The knowledge of such dependencies is useful in various tasks in natural language processing, especially in analysis of sentences involving multiple prepositional phrases, such as: [...] (S-10, 9605013)
4. SHOW: OWN GOAL/PROBLEM IS HARD
   Correctly determining number is a difficult problem when translating from Japanese to English. (S-0, 9511001)
5. DESCRIBE: GENERAL PROBLEM
   The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities. (S-4, 9408011)
6. DESCRIBE: GENERAL CONCLUSION/CLAIM
   It has often been stated that discourse is an inherently collaborative process [...] (S-171, 9504007)
7. DESCRIBE: OTHER CONCLUSION/CLAIM
   Nonetheless there is psychological evidence that language has an unplanned, spontaneous aspect as well (Ochs 1979). (S-9, 9410032)
8. DESCRIBE: OWN GOAL/PROBLEM
   The aim of this paper is to examine the role that training plays in the tagging process [...] (S-32, 9410012)
9. DESCRIBE: OWN CONCLUSION/CLAIM
   [...] *we found that three types of utterances (prompts, repetitions and summaries) were consistently used to signal control shifts.* (S-139, 9504006)
10. DESCRIBE: ARTICLE STRUCTURE
    This paper is organized as follows: We first review a general algorithm for least-errors recognition [...] (S-27, 9502024)
11. PREVIEW: SECTION CONTENTS
    In this section, we are going to motivate the reasons which lead us to choose grammatical words as discriminant. (S-21, 9502039)
12. SUMMARIZE: SECTION CONTENTS
    The previous section provided illustrative examples, demonstrating the performance of the algorithm on some interesting cases. (S-125, 9511006)

Figure 3.7: Moves Based on Swales' CARS Model

ture, our moves 10, 11, 12) are important, even though they have no direct connection to the argumentation. When reporting their research, the authors have to solve the problem of how to linearize their statements in such a way that a reader will be able to understand the main points. In disciplines where fixed section structure is not typical, authors often inform the reader explicitly of which content to expect in each section.

Swales' moves 2A through 2D, which have to do with how other work is introduced and cited, are not included in these 12 moves. In order to operationalize these moves, we should take a closer look at how authors express a stance towards other work, and how this information could be encoded.

3.2.2. Citations and Author Stance

This section will look at results from Content Citation Analysis, one strand of research within library science and the sociology of science, in order to define the concept of authors' stance towards other work. Researchers in content citation analysis have determined and classified semantic relationships between citing and cited works. As we will see it is a highly political matter whether a researcher cites another or not, and what they write about the other's work.

Whereas in industry, the patent system registers intellectual property and thus encourages researchers to produce and contribute new ideas and results, the reward system in science is based on publication and citation (Luukkonen, 1992). To publish an idea means staking a claim of intellectual ownership for that idea (Myers, 1992). The assumption is that other researchers who use the idea must acknowledge them as the authors' intellectual ownership; this is done by formal citation.

Research institutions are rewarded by exercises like the British RAE (Research Assessment Exercise), which measures intellectual output by number of publications in quality journals; individual researchers are affected because publishing is one of the main criteria used in promotion and tenure decisions—this is captured in the well-known motto of "publish or perish".

Other bibliometric measures assesses the quality of a researcher's output, also in a purely quantitative manner, by counting how many papers *cite* a given paper. Content citation analysis is critical of the application of pure citation counting as a measurement of quality and impact of scientific work. Bonzi (1982), for example, points out that *negational* citations, while pointing to the fact that a given work has been *noticed* in a field, does not mean that that work is *received well*, and Ziman (1968), following a slightly different argumentation, states that many citations are done out of “politeness” (towards powerful rival approaches), “policy” (by name-dropping and argument by authority) or “piety” (towards one’s friends, collaborators and superiors). Researchers also often follow the custom of citing some particular early, basic paper, which gives the foundation of their current subject (“paying homage to pioneers”).

Researchers in content citation analysis believe that the classification of motivations is a central element in understanding the relevance of the paper in the field. Many classification schemes for properties of citations have been invented to this end (Weinstock, 1971; Swales, 1990; Oppenheim and Renn, 1978; Frost, 1979; Chubin and Moitra, 1975). Based on such annotation schemes and hand-analyzed data, different influences on citation behaviour can be determined. As one of the earliest such studies, Moravcsik and Murugesan (1975) divide citations in running text into four dimensions: conceptual or operational use (i.e., use of theory vs. use of technical method); evolutionary or juxtapositional (i.e., own work is based on the cited work vs. own work is an alternative to it); organic or perfunctory (i.e., work is crucially needed for understanding of citing article or just a general acknowledgement); and finally confirmative vs. negational (i.e., is the correctness of the findings disputed?). They found, for example, that 40% of the citations were perfunctory, which casts further doubt on the mere citation-counting approach.

As another example of a finer-grained scheme, we reproduce Spiegel-Rüsing’s (1977) scheme (taken from p. 105) in figure 3.8. Spiegel-Rüsing’s results are that of 2309 citations examined, 80% substantiated statements (category 8), 6% discussed history or state of the art of the research area (category 1) and 5% cited comparative data (category 5).

Annotation schemes such as the ones discussed above are subjective, the suggested classifications are difficult to operationalize and annotation is usually not confirmed by reliability studies. Swales (1986), for example, calls researchers in Content Citation Analysis “zealously interpretative” (p. 44).

We are interested in the role that authors’ stance plays in the overall argumentation of the paper, as this stance can provide the information of *relatedness* (e.g. rivalry and ancestry) between papers. It is natural to expect that authors should express a stance towards work they introduce: real estate in the paper is sparse, so authors will tend to try and put it to good use for strengthening the argument. If the other work is used as part of her solution, we expect the author to express a positive stance; if she compares her own work with it or if she has identified a problem with it, we expect a

1. Cited source is mentioned in the introduction or discussion as part of the history and state of the art of the research question under investigation.
2. Cited source is the specific point of departure for the research question investigated.
3. Cited source contains the concepts, definitions, interpretations used (and pertaining to the discipline of the citing article).
4. Cited source contains the data (pertaining to the discipline of the citing article) which are used sporadically in the article.
5. Cited source contains the data (pertaining to the discipline of the citing particle) which are used for comparative purposes, in tables and statistics.
6. Cited source contains data and material (from other disciplines than citing article) which is used sporadically in the citing text, in tables or statistics.
7. Cited source contains the method used.
8. Cited source substantiated a statement or assumption, or points to further information.
9. Cited source is positively evaluated.
10. Cited source is negatively evaluated.
11. Results of citing article prove, verify, substantiate the data or interpretation of cited source.
12. Results of citing article disprove, put into question the data as interpretation of cited source.
13. Results of citing article furnish a new interpretation/explanation to the data of the cited source.

Figure 3.8: Spiegel-Rüsing's (1977) Categories for Citation Motivations

contrastive stance. We also expect other work which is more relevant to receive more space in the paper. While we do not deny that there are many other motivations for citing apart (e.g. citations for general reference, background material, homage to pioneers (Ziman, 1968)), we still assume here that citations which are afforded some space in the paper will be used to support the overall scientific argumentation.

In this context it is interesting to consider negational citations. Both Moravcsik and Murugesan and Spiegel-Rüsing found that negational citations are rare.

MacRoberts and MacRoberts (1984) argue that the reason why pure negational citations are rare is that they are potentially politically dangerous, and that they must therefore be made more acceptable. They claim that authors dissemble in order to diffuse the impact of negative references, hiding a negative point behind insincere praise,