or diffusing the thrust of criticism with perfunctory remarks ("damning them with faint praise"). Brooks's (1986) interviews of scholars and classification of 437 references confirms this hypothesis. In our data we found ample evidence of this effect, cf. the following examples:

> *This account makes reasonably good empirical predictions, though it does fail for the following examples:* (S-75, 9503014)

> *Hidden Markov Models (HMMs) (Huang et al. 1990) offer a powerful statistical approach to this problem, though it is unclear how they could be used to recognise the units of interest to phonologists.* (S-24, 9410022

> *Even though these approaches often accomplish considerable improvements with respect to efficiency or termination behavior, it remains unclear how these optimizations relate to each other and what comprises the logic behind these specialized forms of filtering.* (S-21, 9604019)

When there was apparent simultaneous positive and negative evaluation of a citation in one paper, the positive negation always precedes the negative one, suggesting that the real intention was to criticize.

The moves given in figure 3.9 are based on author stance. The first of these moves describes a weakness of previous research (cf. Spiegel-Rüsing's 10, 12, possibly 13; Moravcsik/Murugesan's "negational/juxtapositional"). The next three describe comparisons between own and other work (cf. Spiegel-Rüsing's category 5; no Moravcsik/Murugesan category). The move expressing the fact that other work is advantageous is best expressed with Spiegel-Rüsing's category 9, and Moravcsik/Murugesan's "confirmative". The final move, a statement of intellectual ancestry, is expressed in many of Spiegel-Rüsing's categories (2, 3, 4, 5, 6, 7, possibly 9), and in Moravcsik/Murugesan's "evolutionary" category.

Note that our main distinction into positive/continuing and negative/contrastive stances can be expected to be intuitive: all annotation schemes enumerated here make this distinction, including Shum's (1998) meta-data scheme. Spiegel-Rüsing's and many other schemes, however, typically make finer distinctions.

13. SHOW: OTHER SOLUTION IS FLAWED
Goal-freezing [...] *is equally unappealing: goal-freezing is computationally expensive, it demands the procedural annotation of an otherwise declarative grammar specification, and it presupposes that a grammar writer possesses substantial computational processing expertise.* (S-59, 9502005)

14. SHOW: OWN SOLUTION IS DIFFERENT FROM OTHER SOLUTION
The use of the chart to store known results and failures allows the user to develop hybrid parsing techniques, rather than relying on the default depth-first top-down strategy given by analysing with respect to the top-most category. (S-146, 9408006)

15. SHOW: OWN GOAL/PROBLEM IS DIFFERENT FROM OTHER GOAL/PROBLEM
Unlike most research in pragmatics that focuses on certain types of presuppositions or implicatures, we provide a global framework in which one can express all these types of pragmatic inferences. (S-124, 9504017)

16. SHOW: OWN CLAIM IS DIFFERENT FROM OTHER CLAIM
Despite the hypothesis that the free word order of German leads to poor performance of low order HMM taggers when compared with a language like English, we have shown that the overall results for German are very much along the lines of comparable implementations for English, if not better. (S-117, 9502038)

17. SHOW: OTHER SOLUTION IS ADVANTAGEOUS
CUG (Categorial Unification Grammar; Uszkoreit (1986)) is advantageous, compared to other phrase structure grammars, for parallel architecture, because we can regard categories as functional types and we can represent grammar rules locally. (S-10, 9411021)

18. STATE: OTHER SOLUTION PROVIDES BASIS FOR OWN SOLUTION
We present a different method that takes as starting point the back-off scheme of Katz (1987). (S-24, 9405001)

Figure 3.9: Moves Based on Author Stance

Our move 18 STATE: OTHER SOLUTION PROVIDES BASIS FOR OWN SOLUTION might well be split into a) theoretical basis b) use of data or c) definition of used methodology—however, what interests us here is the *positive* tenet and the idea of intellectual ancestry more than the exact aspect of agreement with the prior work.

Content citation analysis experiments seem to point to the fact that humans are in principle capable of determining author stance in running text—we will, in section 4.3, employ human judgement for a similar task. However, as already mentioned in section 2.1.2, we are concerned about the potentially high level of subjectivity, a general problem with many studies in the field of content citation analysis.

We try to increase the objectivity of the task by giving exact guidelines and instructing our annotators to only mark citation stance when the authors have explicitly stated it. Also, the most subjective categories are not part of our scheme ("paying homage to pioneers"), which should put us on fairly objective ground. Nevertheless, in order to make sure that these decisions can indeed be made reliably, we also measure reproducibility and stability between several annotators formally.

Other content citation analysis research which is important for us concentrates on relating textual spans to authors' descriptions of other work. For example, in O'Connor's (1982) experiment, *citing statements* (one or more sentences referring to other researchers' work) were manually identified. The main problem encountered in that work is the fact that many instances of citation context are linguistically unmarked. Our data confirms this: articles often contain large segments, particularly in the central parts, which describe research in a fairly neutral way. In order to capture the role of these long neutral segments for the overall argumentation, we needed to define different types of moves. The basis of this definition will be the attribution of intellectual ownership, as motivated in the next section.

3.2.3. Attribution of Intellectual Ownership

We have discussed in the previous section how knowledge claims of *other* authors are acknowledged in the reward system of science. Of course, it is equally essential that the knowledge claims of the current paper itself are registered properly (Myers, 1992), as the intellectual rights to the solution or claim associated with the research are not owned by the authors until they have been accepted by the community via peer review (Zuckerman and Merton, 1973).

Whereas it is arguably in the interest of every researcher to publish as many articles as possible, new research results are a scarce and valuable substance. Research might be presented and possibly perceived as coming naturally in different "sizes"—journal-article-length, conference-length or workshop-length packets of scientific knowledge—but it is clear that this is not how research is done. It is more typically a continuous activity carried out over decades by an individual and her co-workers, such that it is not obvious how much of it should be reported in one paper. Instead, the amount of new research going into a paper is a strategic decision for every researcher.

One strategy for publishing more is to present as many aspects of one piece of research in as many publications as will get accepted, with as few changes as possible. This results in authors breaking research down into "smallest publishable units". This phenomenon is illustrated by clusters of papers with titles which are close variations of one theme—it can be assumed that the scientific innovations presented in these papers will show a high level of overlap. However, there is a tension between the interest of the individual to publish and the interest of the field not to be swamped by near-identical papers. The main quality control mechanism in science is the peer-reviewing process, which guarantees a minimum size of the smallest publishable unit, by making sure that in principle each published paper contains at least *something* new ("original" and "previously unpublished").

A scientific paper contains many ideas and statements which are not the authors' own ideas and beliefs, but which are needed to guide the reader towards accepting their own ideas and beliefs. Other ideas, methods or results are associated with other researchers, namely those which own the intellectual rights for them. Of course, the author does not claim intellectual ownership of those statements; instead, she should recognize the other authors' knowledge claims for them.

We think of documents as divided into segments of different intellectual ownership, where each segment plays a certain role in the overall scientific argumentation:

- General statements about the field's problems and methodologies; statements are portrayed as generally accepted in the field (BACKGROUND).
- More specific descriptions of other researchers' work, e.g. rival approaches (OTHER).
- As the real interest of an author is to stake a new knowledge claim, she needs to make clear what exactly her new contribution is (OWN).

The logical tri-section into types of intellectual ownership is related to the semantics of all moves introduced so far, and it also defines the three new moves shown in figure 3.10. These moves constitute larger textual units than the moves introduced so far which are typically associated with single sentences. For a coverage of the entire paper, the longer moves are indispensable.

We believe that clear attribution of intellectual ownership is one aspect of overall writing quality of a paper: readers often have difficulty recognizing attribution of intellectual ownership in unclearly written papers. Section 4.3.2 will address this question by first experimentally testing if humans can in principle attribute ownership reli-

19. DESCRIBE: GENERAL SOLUTION
The traditional approach has been to plot isoglosses, delineating regions where the same word is used for the same concept. (S-3, 9503002)

20. DESCRIBE: OTHER SOLUTION
Instead, Katz's back-off scheme redistributes the free probability mass non-uniformly in proportion to the frequency of <EQN/>, by setting <EQN/> (S-56, 9405001)

21. DESCRIBE: OWN SOLUTION
The basic idea [...] *is to move from dealing with a single model to dealing with a collection of models linked by an accessibility relation.* (S-196, 9503005)

Figure 3.10: Moves Based on Intellectual Ownership

ably; it will then argue that those texts where they disagree much more than expected must be less clearly written.

How do humans understand who a certain statement in a scientific article is attributed to?

- *Top-down information:* Readers anticipate certain argumentative moves; when interpreting the text they infer the probable communicative intentions of the author.

- *World-Knowledge:* Experts use world knowledge to infer intellectual ownership. They know which statements in a text are established fact and which are intellectually owned by other researchers, and assume that everything else must be the authors' conjecture or knowledge claim.

- *Agent markers:* Agents (other researchers or the authors) typically appear in ritualized roles—they are often portrayed as rival researchers ("*Chomsky argues that*", *"workers in AI"*), as contributors of supportive research (*"several discourse linguists"*) and as representatives of the general opinion in the field (*"It is a well-known fact that*").

- *Segmentation and boundaries:* However, not every sentence contains agent markers. On the contrary, even in clear and well-written papers, most sentences are unmarked propositions which state facts about the object world. Their status can be inferred from surrounding attribution boundaries. Readers assume that unmarked statements are attributed to the previously explicitly mentioned agents, until a new explicit attribution redefines the status of the next segment, or until an obvious conflict catches the reader's eye.

- *Linguistic Cues:* Readers use linguistic cues like tense and voice and non-linguistic cues like location to check that they are still in the type of segment they expect to be in.

Of course, there are papers which show a less pronounced tri-section of intellectual ownership. Work which is "close" to the authors—particularly previous own or co-authored work, but also work of friends or colleagues of the same institution—is usually treated in the text similarly to how the own work is treated, e.g. it is evaluated more positively than other work cited. In some cases, the authors continue a tradition, i.e., add a small amount of research to own previous work described elsewhere. Often the largest part of such papers describes the *previous* own work in a tenet that might make the reader mistake it for the actual *new* contribution of the given paper, if she does not know the prior paper ("smallest publishable unit"). Attribution might then be ambiguous for large portions of the text, an unclarity which might actually even be in the interest of the author.

However, we consider close work as distinct from the current work: As motivated in chapter 1, our task is to determine each paper's contribution with respect to other papers in order to support searchers in a document retrieval environment. Their choice is bound to be particularly difficult if the papers are by the same authors in a similar time frame. The idea is that it is the knowledge claim of each paper which should provide the selection criterion.

In review or position papers, all intellectual work is at a meta-level (reasoning about research work)—no own "technical" object-level work is performed. Thus, the distinction of own and other work does not really apply. A similar case of meta-level research are evaluation papers, i.e. papers in which one approach (typically, one's own) is formally evaluated on a given task, or several approaches are formally compared (one's own approach typically being one of these).

For now, there is one last piece missing in the argumentational mosaic before we can move on to the overall model. This piece has to do with statements describing research as a sequence of (successful or unsuccessful) problem-solving activities.

3.2.4. Statements about Problem-Solving Processes

There are different descriptions of the internal logic of the scientific research process; some of these are oriented in the hypothesis testing framework (Suppe, 1998). An alternative is to regard scientific papers as reports of a problem-solving activity (Hoey, 1979; Solov'ev, 1981; Jordan, 1984; Zappen, 1983; Trawinski, 1989).

In theoretical sciences, the problem is to find an adequate and explanatory *model* that accounts for the evidence obtained from observing the real world, whereas in experimental sciences, the problem is to find *evidence* for some theory about how the world works. In engineering, *artefacts* are designed which fulfill a certain predefined function. Accordingly, what counts as an acceptable solution is discipline specific.

We describe now a simple view of academic research acts. In this model, one atomic research act is associated with exactly one paper. A situation Sit_0 is perceived as unsatisfactory because problem $Prob_0$ is associated with it. The first step in the research process is the formulation of a research goal $Goal_0$. Problem $Prob_0$ is solved (or at least "*addressed*") by applying a solution $Solu_0$ (a new methodology, or an experiment), which leads to a situation Sit_1. Whereas the problem $Prob_0$ might or might not be already known in the field, the solution $Solu_0$ is always assumed to be new (at the least, the application of the solution in the given problem situation is new). Evaluation measures how well the goal was achieved, i.e., how much the overall situation has improved, by implicitly or explicitly comparing situations Sit_0 and Sit_1. There might be remaining problems $Prob_1$ associated with Sit_1 which are not addressed in the current paper; they are the *limitations of the approach*. They are typically portrayed as less severe than the problems which motivated the research ($Prob_0$).

For the argumentation in the paper, Situation Sit_0 needs to be portrayed as undesirable; to improve Sit_0 is the central motivation of the paper. Alternatively, one could show that Sit_1 is desirable; at the very least, situation Sit_1 should be more desirable than situation Sit_0, even if only because in Sit_1 more knowledge is available.

With respect to knowledge claims, the solution is the single entity which is most proprietary about one problem-solving process; the authors want to be attributed with it. To a lesser degree, the research goal can also be considered as the authors' contribution. In some fields, e.g. in complexity theory, the invention of new problems is itself a research goal which would justify the publication of a paper. Such meta-problems do not fit well with our simple problem-solving model.

Not only can the *own* problem-solving process be described by such atomic research acts. The argumentation in a paper also involves descriptions of other people's problem-solving activities. The background of a problem can be introduced as (possibly successive) problem-solving actions, including general problems in the field, general solutions, research goals and evaluation methodologies. The problem addressed in the paper (*Prob*$_0$) could be a specific weakness of prior solutions which have led to the situation *Sit*$_0$, or it could be a general, long-standing problem in the field.

The own solution can be portrayed as building on some other problem-solving process: some other methodology or idea is taken as the basis for the reported research and applied either with or without changes.

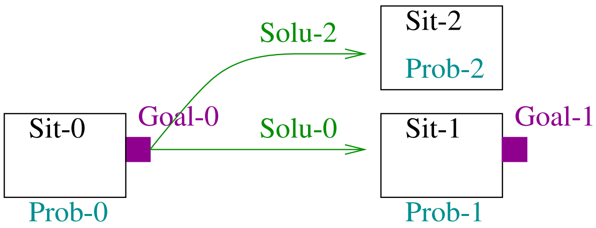


Figure 3.11: Rival Problem-Solving Processes

Figure 3.11 shows a situation where the own paper solution *Solu*$_0$ solves a *known* problem *Prob*$_0$, i.e. a problem to which some other researchers have already presented a solution *Solu*$_2$. The problem solving process presented by the other researchers leads to a different situation *Sit*$_2$. *Sit*$_2$ is similar to *Sit*$_1$, the one favoured by the authors, in that both *Sit*$_1$ and *Sit*$_2$ are not associated with the original problem *Prob*$_0$ anymore, but they differ in some other respect. It is the task of the authors to motivate that the own solution is better than the rival solution. For example, there might be (new) problems associated with *Solu*$_2$, or *Solu*$_2$ might be inferior according to some default criteria—solutions are supposed to be explanatory, elegant, simple, and efficient.

Statements about own and other problem solving processes abound in our data. Figure 3.12 summarizes our moves based on author stance and problem-solving statements. Note that moves describing somebody else's unsuccessful problem solving activity also express contrastive stance and could have been classified as belonging to the moves in figure 3.9.

As the reader has now seen almost all moves we propose and should have an idea of the constructions this thesis is interested in, we will turn to the important aspect of *how* such statements are typically expressed in scientific articles.

22. SHOW: OWN SOLUTION SOLVES OWN PROBLEM
This account also explains similar differences in felicity for other coordinating conjunctions as discussed in Kehler (1994a) [...] (S-100, 9405010)

23. SHOW: OWN SOLUTION IS NECESSARY TO ACHIEVE OWN GOAL
We have argued that obligations play an important role in accounting for the interactions in dialog. (S-217, 9407011)

24. SHOW: OWN SOLUTION AVOIDS PROBLEM
This paper presents a treatment of ellipsis which avoids these difficulties, while having essentially the same coverage as Dalrymple et al. (S-9, 9502014)

25. SHOW: OTHER SOLUTION DOES NOT SOLVE PROBLEM
Computational approaches fail to account for the cancellation of pragmatic inferences: once presuppositions or implicatures are generated, they can never be cancelled. (S-20, 9504017)

26. SHOW: OTHER SOLUTION SOLVES PROBLEM
The Direct Inversion Approach (DIA) of Minnen et al. (1995) overcomes these problems by making the reordering process more goal-directed and developing a reformulation technique that allows the successful treatment of rules which exhibit head-recursion. (S-15, 9502005)

27. SHOW: OTHER SOLUTION INTRODUCES NEW PROBLEM
Specifically, if a treatment such as Hinrichs's is used to explain the forward progression of time in example <CREF/>, then it must be explained why sentence <CREF/> is as felicitous as sentence <CREF/>. (S-12, 9405002)

28. SHOW: OWN SOLUTION IS BETTER THAN OTHER SOLUTION
We found that the MDL-based method performs better than the MLE-based method. (S-11, 9605014)

29. SHOW: OWN GOAL/PROBLEM IS HARDER THAN OTHER GOAL/PROBLEM
[...] *disambiguating word senses to the level of fine-grainedness found in WordNet is quite a bit more difficult than disambiguation to the level of homographs (Hearst 1991; Cowie et al. 1992).* (S-147, 9511006)

Figure 3.12: Moves Based on Problem-Solving Statements

3.2.5. Scientific Meta-Discourse

In section 3.2.3 we hypothesized that there are superficially recognizable correlations of boundaries of zones of intellectual attribution, e.g. expressions like *"Chomsky claims that"*. We believe that meta-discourse is one of the most universally applicable structure markers in scientific text.

Meta-discourse, commonly defined as *discourse about discourse*, is a name for

Category	Function	Examples
Textual meta-discourse		
Logical connectives	express semantic relation between main clauses	*in addition; but; therefore; thus*
Frame markers	refer to discourse acts or text stages	*to repeat;our aim here; finally*
Endophoric markers	refer to information in other parts of the text	*noted above; see Fig 1; below*
Evidentials	refer to source of information from other texts	*according to X; Y (1990)*
Code glosses	help readers grasp meanings of ideational material	*namely; eg; in other words*
Interpersonal meta-discourse		
Hedges	withhold author's full commitment to statements	*might; perhaps; it is possible*
Emphatics	emphasize force or author's certainty in message	*in fact; definitely; it is clear; obvious*
Attitude markers	express author's attitude to propositional content	*surprisingly; I agree; X claims*
Relational markers	explicitly refer to or build relationship with reader	*frankly; note that; you can see*
Person markers	explicit reference to author(s)	*I; we; my; mine; our*

Figure 3.13: Hyland's (1998) Categories of Meta-Discourse

all those statements which fulfill other functions but to convey pure propositional contents (the "science" in the paper). Meta-discourse is a pragmatic construct by which writers signal their communicative intentions (Hyland, 1998; Swales, 1990). It is ubiquitous in scientific writing: Hyland (1998) found a meta-discourse phrase on average after every 15 words in running text, hedges being the most frequent type of meta-discourse in his texts. His classification of meta-discourse is given in figure 3.13.

Some of Hyland's categories (Attitude markers, Person markers, Evidentials, Endophorics and Frame Markers) seem immediately relevant to the effects discussed in this chapter. Another set of meta-discourse which we are particularly interested in are meta-statements about the own research. Much of that type of scientific meta-discourse is conventionalized, particularly in experimental sciences, and particularly in the methodology or result section; linguistically, there is not much variation (e.g. "*we present original work...*", or "*An ANOVA analysis revealed a marginal interac-*

tion/a main effect of..."). Such formulaic expressions occur less often in the discussion section and the introduction where there is more room for personal style. Swales (1990) lists many such fixed phrases as co-occurring with the moves of his CARS model (p.144;pp.154–158;pp.160–161). Another type of meta-discourse points to the current research process ("*in this paper*", "*here*"), expresses affect ("*unfortunately*") or knowledge states ("*to the best of our knowledge*"; "*it has long been known*").

It is well-known that different disciplines use different meta-discourse. Hyland (1998) argues that meta-discourse variation between scientific communities can be attributed to the fact that meta-discourse has to follow the norms and expectations of particular cultural and professional communities—scientific communities impose linguistic standardization pressures. He found significant differences in meta-discourse use across disciplines (Microbiology, Marketing, Astrophysics and Applied Linguistics), though the articles displayed a remarkable similarity in the *density* of meta-discourse. Marketing and Applied Linguistics papers used far more interpersonal meta-discourse than those in Biology and Astrophysics, which, on the other hand, use far more textual meta-discourse. Due to the particularities of our data we expect meta-discourse in our corpus to be varied.

And even within one discipline, there is a large class of expressions which express similar, prototypical moves, even though the resulting sentences do not look similar on the surface. This is particularly the case for statements referring to aspects of the problem-solving process or to the author's stance towards other work: expressions of contrast to other researchers and for statements of research continuation. Figure 3.14 shows that there are many ways to express the fact that one piece of work is based on some previous other work.

The surface forms of these sentences are very different despite the similar semantics they express: in some sentences the syntactic subject is a method, in others it is the authors, and in others the originators of the based-upon idea. Also, the verbs used are very different. This wide range of linguistic expression presents a real challenge—later parts of this thesis will be concerned with finding a method for recognizing a large subset of such variable meta-discourse (cf. section 5.2.2).

After this brief look at the syntactic variability of the moves, we now return to our model of overall strategy of argumentation.

• *Thus, we base our model on the work of Clark and Wilkes-Gibbs (1986), and Heeman and Hirst (1992) who both modeled (the first psychologically, and the second computationally) how people collaborate on reference to objects for which they have mutual knowledge.* (S-15, 9405013)

• *The starting point for this work was Scha and Polanyi's discourse grammar (Scha and Polanyi 1988; Pruest et al. 1994).* (S-4, 9502018)

• *We use the framework for the allocation and transfer of control of Whittaker and Stenton (1988).* (S-36, 9504007)

• *Following Laur (1993), we consider simple prepositions (like "in") as well as prepositional phrases (like "in front of").* (S-48, 9503007)

• *Our lexicon is based on a finite-state transducer lexicon (Karttunen et al. 1992).* (S-2, 9503004)

• *Instead of feature based syntax trees and first-order logical forms we will adopt a simpler, monostratal representation that is more closely related to those found in dependency grammars (e.g. Hudson (1984)).* (S-116, 9408014)

• *The centering algorithm as defined by Brennan et al. (BNF algorithm), is derived from a set of rules and constraints put forth by Grosz et al. (Grosz et al. 1983; Grosz et al. 1986).* (S-56, 9410006)

• *We employ Suzuki's algorithm to learn case frame patterns as dendroid distributions.* (S-23, 9605013)

• *Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition.* (S-151, 9405001)

Figure 3.14: Variability of Statements Expressing Research Continuation

3.2.6. Strategies of Scientific Argumentation

Scientific articles are *biased* reports; the argumentation follows the interest of the author. Indeed, we see the whole paper as one rhetorical act, as Myers (1992) does. The high level communicative goal in a paper, apart from conveying a message, is to persuade the scientific community of the relevance, reliability, quality and importance of the work (Swales, 1990; Kircz, 1998). There are parallels to politeness theory (Brown and C., 1987), where the commodity that is traded is "face"; in the case of scientific writing, the commodity is "credibility".

There are some "high level" moves which are essential for the overall argumentation: One needs to show that the research process is successful, i.e. that the total knowledge available to the community must have increased. The most important ones

SHOW: OWN RESEARCH IS VALID CONTRIBUTION TO SCIENCE

SHOW: RESEARCH IS JUSTIFIED

SHOW: AUTHORS ARE KNOWLEDGEABLE

SHOW: OTHER RESEARCHERS HAVE TRIED TO SOLVE THE PROBLEM

SHOW: OWN SOLUTION PROCESS IS NEW

SHOW: NOBODY HAS USED SAME SOLUTION FOR SAME PROBLEM BEFORE

30. SHOW: OWN GOAL/PROBLEM IS NEW
[...] *and to my knowledge, no previous work has proposed any principles for when to include optional information* [...] (S-9, 9503018)

31. SHOW: OWN SOLUTION IS ADVANTAGEOUS
The substitutional treatment of ellipsis presented here [...] *has the computational advantages of* [...] (S-210, 9502014)

Figure 3.15: Moves Based on Higher-Level Intentions

of these moves are given in figure 3.15.

The first six moves in figure 3.15 are not numbered and contain no corpus example. The reason for this is that these moves are not typically made explicit; instead, the reader is left to induce them. The last two high-level moves, however, do occur explicitly, making our set of 31 argumentative moves complete (summarized in figure 3.16).

Relations between the moves are shown in figure 3.17. The tree relation means "Is A Sub-Move Of". An argumentation strategy might be as follows: One might say that the own problem is hard, then introduce the own solution, argue that it solves the problem, argue that this solution is better than somebody else's solution or state the fact that the problem has never been addressed before.

Not all of these moves have to occur in a scientific article for the argumentation to be successful or complete. For example, the problem addressed ($Prob_0$) can be new to the field; this can be stated explicitly (30). Additionally, one can shown that similar problems addressed before are different from the given one. This would additionally fulfill the function of showing that the authors are knowledgeable in their field. But problems need not be new; they might have been addressed by others before (cf. the

I. Moves borrowed from Swales

1. DESCRIBE: GENERAL GOAL
2. SHOW: OWN GOAL/PROBLEM IS IMPORTANT/INTERESTING
3. SHOW: SOLUTION TO OWN PROBLEM IS DESIRABLE
4. SHOW: OWN GOAL/PROBLEM IS HARD
5. DESCRIBE: GENERAL PROBLEM
6. DESCRIBE: GENERAL CONCLUSION/CLAIM
7. DESCRIBE: OTHER CONCLUSION/CLAIM
8. DESCRIBE: OWN GOAL/PROBLEM
9. DESCRIBE: OWN CONCLUSION/CLAIM
10. DESCRIBE: ARTICLE STRUCTURE
11. PREVIEW: SECTION CONTENTS
12. SUMMARIZE: SECTION CONTENTS

II. Moves defined by author stance

13. SHOW: OTHER SOLUTION IS FLAWED
14. SHOW: OWN SOLUTION IS DIFFERENT FROM OTHER SOLUTION
15. SHOW: OWN GOAL/PROBLEM IS DIFFERENT FROM OTHER GOAL/PROBLEM
16. SHOW: OWN CLAIM IS DIFFERENT FROM OTHER CLAIM
17. SHOW: OTHER SOLUTION IS ADVANTAGEOUS
18. STATE: OTHER SOLUTION PROVIDES BASIS FOR OWN SOLUTION

III. Moves defined by attribution of ownership

19. DESCRIBE: GENERAL SOLUTION
20. DESCRIBE: OTHER SOLUTION
21. DESCRIBE: OWN SOLUTION

IV. Moves defined by problem solving statements

22. SHOW: OWN SOLUTION SOLVES OWN PROBLEM
23. SHOW: OWN SOLUTION IS NECESSARY TO ACHIEVE OWN GOAL
24. SHOW: OWN SOLUTION AVOIDS PROBLEMS
25. SHOW: OTHER SOLUTION DOES NOT SOLVE PROBLEM/DOES NOT ACHIEVE GOAL
26. SHOW: OTHER SOLUTION SOLVES PROBLEM
27. SHOW: OTHER SOLUTION INTRODUCES NEW PROBLEM
28. SHOW: OWN SOLUTION IS BETTER THAN OTHER SOLUTION
29. SHOW: OWN GOAL/PROBLEM IS HARDER THAN OTHER GOAL/PROBLEM

V. High level moves

30. SHOW: OWN GOAL/PROBLEM IS NEW
31. SHOW: OWN SOLUTION IS ADVANTAGEOUS

Figure 3.16: List of Argumentative Moves

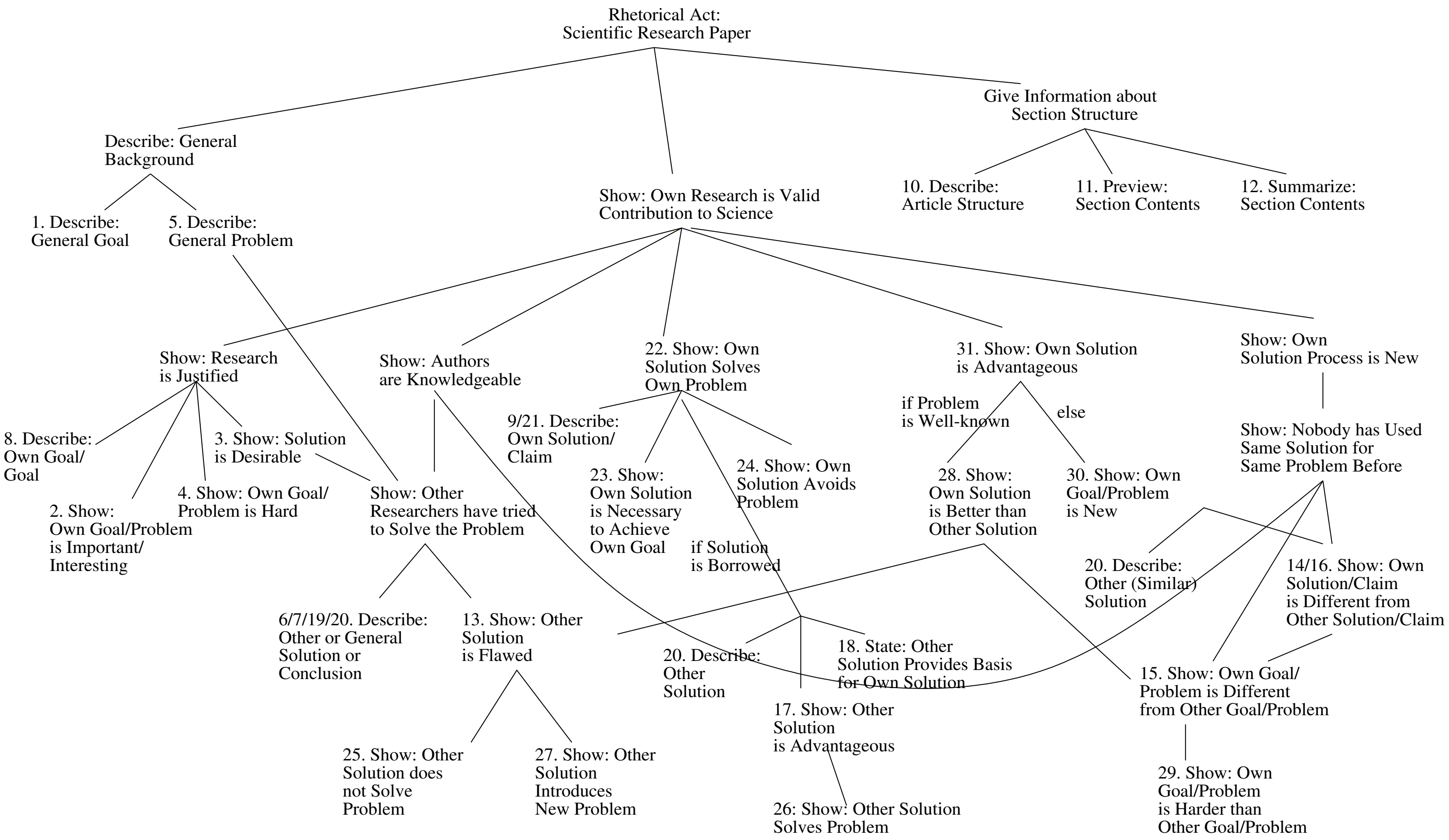


Figure 3.17: Relations between Argumentative Moves

situation in figure 3.11, where a rival solution was suggested). In that case, one needs to show that the own solution is better (28) or that the other solution is flawed (25 or 27).

All of the moves cover a textual span at least as long as a sentence, and in some cases they cover much larger textual spans. Some moves—particularly the moves of type SHOW—can be explicitly stated in one single sentence, but many moves typically span longer segments, for example the moves of type DESCRIBE, which detail problems, solutions and goals in a neutral way and whose purpose is informative rather than rhetorical. We consider the whole move as one unit for our purposes, disregarding possible internal move structure.

Some moves in the diagram tend to occur with other moves, e.g., moves describing other work (6, 7, 19 or 20) co-occur with statements about the role of this other work for the current work (critical stance in moves 13, 25, 27; contrastive stance in moves 14, 15, 16, 29; positive stance in moves 17, 18, 26). Relations of such kinds between moves are not shown in the diagram.

Moves sometimes serve more than one communicative and argumentative purpose at once. The move OTHER RESEARCHERS HAVE TRIED TO SOLVE THE PROBLEM describes the history of the problem, provides background knowledge, proves that the authors know the literature in the field, and it shows that the problem is indeed justified and that a solution is desirable.

3.3. An Annotation Scheme for Argumentative Zones

In the previous section, we have introduced a rather complex model of discourse and argumentative effects in scientific text. We believe that our implicit claim—that the model explains our data adequately—should be substantiated by demonstrating that other humans can apply the account consistently to actual texts. In this section, we will operationalize our model by defining a practical annotation based on it.

In general, designing an annotation scheme has many pitfalls. One wants the annotation scheme to be a) predictive and informative, so that it will prove useful for an end task and b) intuitive, or at least learnable, such that it can be applied consistently by different annotators and over time. If an annotation scheme is simple and intuitive and the task well-described, it will result in high consistency, but there is a danger that the information contained in it might not be informative enough for the given task. On the other hand, if the categories are informative their definition is necessarily vague, leaving a lot of leeway for subjective interpretation. In this case, it is likely that different annotators will disagree in their judgements. The process of finding a workable annotation scheme is thus a tight rope act between the conflicting requirements of informativeness and consistency. This section reports on our quest for a good annotation scheme, and shows why two predecessors of the final annotation scheme fall short of the requirements.

The first annotation scheme (Teufel, 1998) contains 23 categories defined directly by argumentative moves, similar to those in figure 3.16. Such a scheme based on moves is very informative and encodes valuable information for subsequent fact extraction from the sentences. For example, a sentence of type "SHOW: OWN SOLUTION IS ADVANTAGEOUS" contains both a mention of the own solution and a statement of the advantage of the own solution, a fact which could be exploited for information extraction from such a sentence.

We used two unrelated annotators in the definition phase. As is typical for high-level, information-rich classification tasks, the annotation scheme had to be changed repeatedly during this time. Settling on an exhaustive list of moves which annotators agreed on proved very difficult. We were constantly tempted to add more moves for situations where a given sentences does not quite fall into the semantics already defined. Once the scheme mentioned above (23 categories) had emerged, we wrote guidelines detailing criteria for each move.

After the definition phase, we ran a pilot study with our two, by now, task-trained annotators. This experiment revealed that the scheme was not reliable. Even repeated changes to the annotation scheme at this late stage did not improve agreement significantly. Within the mind of one annotator, private understandings of these categories may well be rather consistent—we annotated 10 randomly sampled, previously annotated papers again after 4 weeks and achieved reasonable agreement with the previous annotation (the concept of stability will be introduced in section 4.2). However, if these understandings cannot be communicated to others, something is wrong with the scheme. Low agreement between different annotators (*reproducibility*; detailed in section 4.2) finally convinced us that a fixed, exhaustive list of such high-level categories at this pragmatic level is not universal enough to train annotators.

In order to make the next scheme easier and more objective, we reduced the number of categories and simplified their definitions, while trying to retain as much of the information as possible for our task. Our second attempt at an annotation scheme

B	BACKGROUND
T	TOPIC
W	RELATED WORK
P	PURPOSE/PROBLEM
S	SOLUTION/METHOD
R	RESULT
C	CONCLUSION/CLAIM

Figure 3.18: Annotation Scheme Based on Functional Abstract Units

(figure 3.18) consisted of just seven categories (Teufel and Moens, 1998, 1999a), which are similar to the functional units well-known from summarizing guidelines (cf. section 2.3.1.2).

Again, we achieved respectable stability when re-annotating parts of the corpus. This is a good sign, but we nevertheless noticed fundamental problems with the type of annotation. It proved extremely difficult to associate textual units as big as sentences (i.e. propositional contents) with categories which describe high-level concepts (i.e. nominal phrases). An additional, orthogonal problem was the fact that some high level entities such as PURPOSE/PROBLEM and SOLUTION can be difficult to distinguish in real-world text. To give an example, we were not sure about the right annotation for the following sentence:

> *We then show how different classes of pragmatic inferences can be captured using this formalism, and how our algorithm computes the expected results for a representative class of pragmatic inferences.* (S-29, 9504017)

Is the sentence to be counted as TOPIC, because "*pragmatic inferences*" are the TOPIC of the paper? Or is it rather the case that *"capturing different classes of pragmatic inferences"* is the PROBLEM/PURPOSE? Or should this sentence be classified as SOLUTION, as the phrase *"our algorithm computes the expected results*" could be interpreted as a high level description of the approach used?

Allowing for multiple annotation seemed to ameliorate the problems, but it lead to so many multiply annotated sentences that we started doubting the informativeness contained in this annotation. We redesigned the scheme radically, resulting in the third and final annotation scheme (figure 3.19).

A simpler version of the scheme (the “basic scheme”) encodes only intellectual ownership (figure 3.20). Pilot studies with our annotators with both schemes showed that they were much more comfortable and accurate when applying these schemes to real texts. These are the schemes we will use for the extensive human annotation experiments reported in chapter 4 (Teufel et al., 1999), and for the prototypical implementation reported in chapter 5 (Teufel and Moens, 1999b).

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weakness of other solution
BASIS	Other work provides basis for own work

Figure 3.19: Final Annotation scheme—Full Version

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...

Figure 3.20: Final Annotation Scheme—Basic Version

As with the other annotation schemes, the categories are to be read as mutually exclusive labels, one of which is attributed to each sentence. Each category is associated with a colour to make human annotation more mnemonic.

We call the categories which occur only in the full scheme but not in the basic scheme *non-basic* categories (i.e. AIM, CONTRAST, TEXTUAL and BASIS). The seven

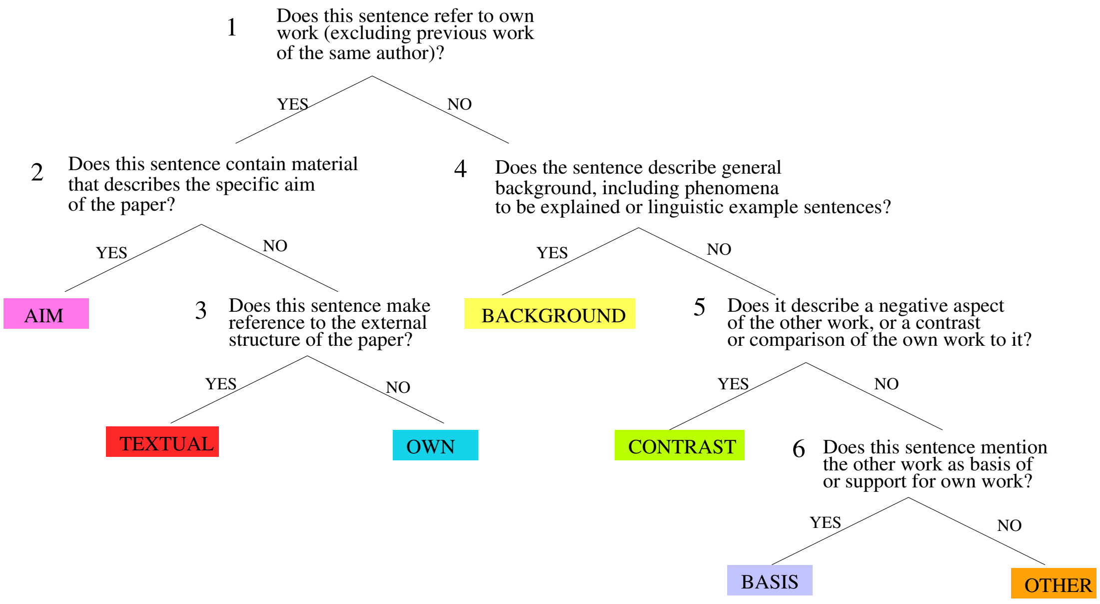


Figure 3.21: Decision Tree for Full Annotation Scheme

categories of the full annotation scheme are closely related to the different aspects of our model (Swales' categories, author stance, intellectual ownership, and problem-solving statements). The semantics of our scheme is best explained with the decision tree in figure 3.21, based on six yes/no questions.

Question 1 focuses on attribution of ownership, distinguishing between statements which describe the authors' own *new* contributions and those which describe research outside the given paper, including the authors' own previous work, generally accepted statements and statements which are attributed to other, specific researchers.

Once annotators decide that the statement describes own work, Question 2 determines AIM sentences. Such sentences describe the research goal addressed in the paper. The most explicit type of AIM sentences is provided by move 8 (DESCRIBE OWN GOAL/PROBLEM in figure 3.16). But dependent on the annotators' intuitions, other moves can in principle be AIM sentences too, e.g. moves 2, 3, 4, 22, 23, 24, 30 and 31.

Question 3 singles out TEXTUAL sentences, i.e. those giving explicit information about section structure. This corresponds to moves 10, 11 and 12. All other statements about own work, in particular move 21, but also all moves not deemed AIM sentences, receive the label OWN.

Question 4 distinguishes between BACKGROUND material (i.e. generally accepted statements; move 1, 5, 6 and 19) and more specifically characterized other work. If the annotators have decided that the sentence describes specific work, then the last two questions concentrate on author stance. Question 5 checks if the other work is presented critically or as problem-wrought (as in moves 13, 25, 27), contrastively (moves 14, 15 and 16), or as inferior to the own solution (moves 28 and 29); in that case, the sentence is assigned to category CONTRAST. Otherwise, Question 6 assigns the category BASIS to statements of research continuation (move 18). Explicit positive statements about other work (i.e. moves 17 and 26) can also be assigned to BASIS. Neutral descriptions of other work get assigned the category OTHER. Details and decision criteria on how to answer the questions are given in the guidelines (cf. appendix C.2).

The relation between the categories and the moves is complex: it is not the case that the categories are super-classes of the moves. Instead, many moves can end up as different zones, depending on the question if there were more appropriate moves to act as argumentative categories. For example, move 3. SHOW: SOLUTION IS DESIRABLE *could* be annotated as AIM in the absence of a move 7; otherwise, it would more appropriately be annotated as OWN. Rather, the seven categories should be seen as a workable compromise between simplicity and informativeness for our document retrieval task.

The task is defined as classification, but it can also be seen as a segmentation task. Because the kind of annotation we envisage includes contiguous, non-overlapping and non-hierarchical sequences, we refer to the segments of sentences with the same category as *zones*. We then call the process of annotation with our argumentative scheme *Argumentative Zoning*. To give an illustration of the task of Argumentative Zoning, figures 3.22 and 3.23 show the first page of our example paper, annotated by us with both versions of the annotation scheme. More human example annotations can be found in the guidelines in the appendix (p. 310, 311, 327 and 328).

Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in 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 probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

BACKGROUND OTHER OWN

Figure 3.22: First Page of Example Paper, Annotated with Basic Annotation Scheme

Distributional Clustering of English Words

Fernando Pereira Naftali Tishby Lillian Lee

Abstract

We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts. Deterministic annealing is used to find lowest distortion sets of clusters. As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to held-out data.

Introduction

Methods for automatically classifying words according to their contexts of use have both scientific and practial interest. The scientific questions arise in connection to distributional views of linguistic (particularly lexical) structure and also in relation to the question of lexical acquisition both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for deciding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example the frequencies of pairs of transitive main verb and the head of its direct object, cannot be reliably used for comparing the likelihoods of different alternative configurations. The problem is that in 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 probabilties.

Hindle (1990) proposed dealing with the sparseness problem by estimating the likelihood of unseen events from that of "similar" events that have been seen. For instance, one may estimate the likelihood of a particular direct object for a verb from the likelihoods of that direct object for similar verbs. This requires a reasonable definition of verb similarity and a similarity estimation method. In Hindle's proposal, words are similar if we have strong statistical evidence that they tend to participate in the same events. His notion of similarity seems to agree with our intuitions in many cases, but it is not clear how it can be used directly to construct classes and corresponding models of association.

Our research addresses some of the same questions and uses similar raw data, but we investigate how to factor word association tendencies into associations of words to certain hidden senses classes and associations between the classes themselves. While it may be worthwhile to base such a model on preexisting sense classes (Resnik, 1992), in the work described here we look at how to derive the classes directly from distributional data. More specifically, we model senses as probabilistic concepts or clusters c with corresponding cluster membership probabilities <EQN/> for each word w. Most other class-based modeling techniques for natural language rely instead on "hard" Boolean classes (Brown et al., 1990). Class construction is then combinatorically very demanding and depends on frequency counts for joint events involving particular words, a potentially unreliable source of information, as we noted above. Our approach avoids both problems.

Problem Setting

In what follows, we will consider two major word classes, <EQN/> and <EQN/>, for the verbs and nouns in our experiments, and a single relation between a transitive main verb and the head noun of its direct object. Our raw knowledge about the relation consists of the frequencies <EQN/> of occurrence of particular pairs <EQN/> in the required configuration in a training corpus. Some form of text analysis is required to collect such a collection of pairs. The corpus used in our first experiment was derived from newswire text automatically parsed by Hindle's parser Fidditch (Hindle, 1993). More recently, we have constructed similar tables with the help of a statistical part-of-speech tagger (Church, 1988) and of tools for regular expression pattern matching on tagged corpora (Yarowsky, p.c.). We have not yet compared the accuracy and coverage of the two methods, or what systematic biases they might introduce, although we took care to filter out certain systematic errors, for instance the misparsing of the subject of a complement clause as the direct object of a main verb for report verbs like "say".

We will consider here only the problem of classifying nouns according to their distribution as direct objects of verbs; the converse problem is formally similar. More generally, the theoretical basis for our method supports the use of clustering to build models for any n-ary relation in terms of associations between elements in each coordinate and appropriate hidden units (cluster controids) and associations between these hidden units.

BACKGROUND OTHER OWN AIM CONTRAST BASIS TEXTUAL

Figure 3.23: First Page of Example Paper, Annotated with Full Annotation Scheme

3.4. Argumentative Zones and RDP Slots

In this thesis, we originally set out to generate RDPs. The semantics of the individual argumentative zones are obviously very close to RDP slots, but argumentative zones and RDP slots are not the same. We will now discuss the relation between the two.

Argumentative zones can be seen as providing the material of text which *might* go into the RDP slots. In a subsequent processing step not treated in this thesis, full RDPs could be created from the information contained in argumentative zones. The RDP presented in section 2.3.2 was manually created based on an annotated version of the example paper, obtained in the annotation exercise to be described in chapter 4.

Some of the zones, the non-basic categories, are short and contain important information; they can therefore act as *direct slot fillers* without requiring much further work. AIM zones, for example, constitute a good characterization of the entire paper, which is typically only one sentence long. They are thus already extremely useful for the generation of abstracts.

But BACKGROUND, OTHER and OWN are longer zones, which should be seen as search ground for later processes. For example, as simple sentence extraction does not take the context of a sentence into account, a selected sentence might turn out to be describing *other* people's work. This is a grave error, particularly if the sentence expresses a statement which the authors reject. By searching and extracting from argumentatively zoned articles, where zones such as OWN and OTHER are distinguished, this error should be eliminated.

There is another task which argumentative zones as search ground is useful for. This task is the association of identifiers of other work (formal citations, names of researchers, names of solutions) with the statement that expresses the author's stance towards the work.

This task is needed in order to generate RDPs from argumentative zones. Our approach has a more concise definition of citation context (cf. O'Connor's (1982) work) than previous approaches. Citation maps display only *one* sentence, namely the sentence which expresses the evaluative statement. In contrast, Lawrence et al.'s (1999) CiteSeer (which displays contexts in a text extract fashion, cf. the example on p. 34), and Nanba and Okumura's (1999) tool operate with a much *larger* citation context. Consider Nanba and Okumura's example of a contrastive citation context (taken from p. 927):

1 In addition, when Japanese is translated into English, the selection of appropriate determiners is problematic.
2 Various solutions to the problems of generating articles and possessive pronouns and determining countability and number have been proposed **[Murata and Nagao, 1993]**.
3 The difference between the way numerical expressions are realized in Japanese and English has been less studied.
4 In this paper we propose an analysis of classifiers based on properties in both Japanese and English.
5 Our category of classifier includes both Japanese josushi 'numerical classifiers' and English partitive nouns.

Nanba and Okumura's tool displays sentences **2–4** (the *reference area*). In our approach, only sentence **3** would be displayed, which implies that one must additionally determine which other work the current context refers to. In this case, the formal citation in sentence **2** must be extracted. As an additional difficulty, the authors might have used different kinds of identification of the other work, e.g. author name or solution identifier. We aim to treat these types of identification alike, instead of recognizing only formal citations (like Nanba and Okumura do).

Nanba and Okumura's approach relies on the simplifying assumption that identification and citation of an approach occur in the same sentence, or at least very close together. However, this does not have to be the case. In our example paper, the description of the work of Hindle (1993) and its weaknesses extends from sentences 5 to 9. Textual separation is an issue that needs to be addressed, as it is even *more* likely for important references, where the authors will take some time and space describing the other work (we also noticed that textual separation is more likely for CONTRAST zones than for BASIS zones, as these are often longer).

Argumentative zones can help us associate textual spans belonging to authors' descriptions of other work because of regularities between zones which we call rhetorical patterns. For example, neutral descriptions of other researchers' work often occur in combinations with statements expressing a stance towards that work. We believe that those kinds of dependencies can be helpful for automatic Argumentative Zoning: in section 5.3.4.2, we will use an ngram model operating over sentences to model these regularities. From informal inspections of our corpus, however, we suspect that in our corpus the dependencies are not as strong as Swales' claims about fixed order would imply—possibly due to the interdisciplinarity of our corpus.

Figure 3.24 illustrates typical argumentative patterns. The identifiers (i.e. re-

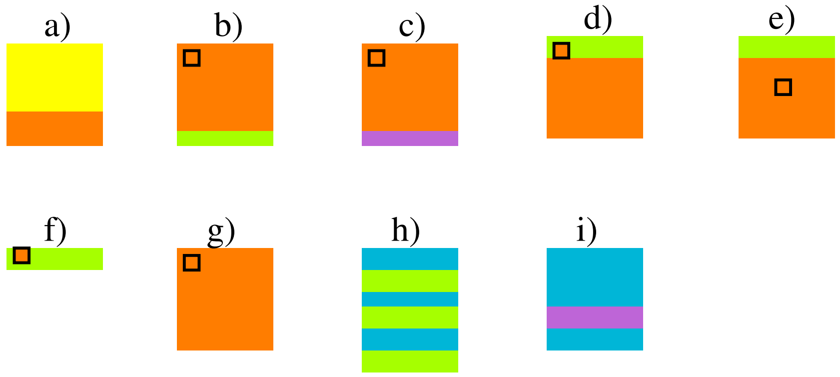


Figure 3.24: Typical Rhetorical Patterns

searchers' names, formal citations or solution names) are signified by small squares.

a) General statements typically precede more specific ones; e.g., general background material is followed by descriptions of specific other work.

b) A prototypical pattern for CONTRAST: The other solution is identified, described and criticized.

c) A prototypical pattern for BASIS: The other solution is identified and described, then a statement of intellectual ancestry follows.

d) The other work is identified and criticized before it is described. This pattern is rather common, though it does not occur as frequently as pattern b).

e) The other work is identified after it has been described and criticized. This pattern reads somewhat awkwardly, but it does occur several times in our corpus.

f) A less important contrastive approach which does not get much "real estate" in the paper.

g) Other work is introduced and identified, but no stance is expressed. In section 3.2.2 we argued that such patterns contribute nothing to the argumentation and that the authors waste space in the paper which such moves. Nevertheless, we found many such patterns in our corpus. One of the possible reason why they were were used nevertheless is that they serve the move SHOW: AUTHORS ARE KNOWLEDGEABLE. As predicted, most of patterns g) found in our corpus are short, i.e. the work is presumably not crucial to the argumentation.

h) After the own solution has been introduced, advantages of it can be presented by comparisons to other work, often in parallel steps. This is a prototypical pattern for comparisons with other work, particularly in conclusions and discussion sections.

i) A statement of intellectual ancestry occurs in the middle of a description of the own solution. We found that if some other work is cited in an OWN segment, it is generally more likely to be a BASIS zone than a CONTRAST zone. BASIS zones are also overall shorter than CONTRAST zones; many of these statements just state the fact that work is based on other work, or acknowledge methods or data used.

In an approach based on Argumentative Zoning, adjacency of argumentative zones and assumptions about their connection to a given zone can be used to find the most likely citation association. For example, if a zone expressing author stance has been identified which does not contain an identifier, adjacent zones of other researchers' work can be searched for identifiers most likely to be associated with the zone.

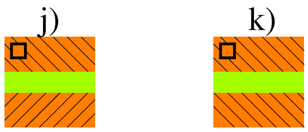


Figure 3.25: Likely and Unlikely Rhetorical Patterns

An aide in this could be provided by the following observation which is illustrated in figure 3.25: we found that if two zones of neutral description occur around a criticism zone, it is very unlikely that the neutral zones refer to the same work (as in j); it is far more likely that they refer to different work (as in k).

Additionally, argumentative zones could be used in Content Citation analysis to provide a simple and automatic means of estimating the importance of a cited work for the citing work, as more relevant OTHER work will probably receive more space in the article.

3.5. Related Work

Argumentative Zoning is a new task, but there is much work in computational and theoretical linguistics and in language engineering which is closely related. Firstly, there are other types of zoning of text, i.e. methods which break documents into segments; it is the definition of the zones which is new in our approach. While most other approaches try to segment papers into topic-related zones (Morris and Hirst, 1991; Hearst, 1997), our approach is more similar in nature to Wiebe's (1994) work. Her approach also attempts to determine a rhetorical feature, namely *evidentiality* or point of view in narrative. The task is to determine the source of information in text which might be either subjective or objective. In news reporting and narrative, this distinction is important as coherent segments presenting opinions and verbal reactions are mixed with segments presenting objective fact. Her four categories are given in figure 3.26 (examples taken from Wiebe et al. 1999, p. 247).

Subjectivity is a property which is related to the attribution of authorship as well as to author stance, but there are obvious differences between Wiebe's and our distinction, which are rooted in differences between the text types covered. As will be discussed in chapter 5, some of the sentential features we use are comparable to hers (e.g. occurrence of first or third person personal pronouns). However, her processing does not go as "deep" as ours in trying to determine the agent/action structure of the text.

Another kind of discourse segment altogether is defined by topic segments (Morris and Hirst, 1991; Kozima, 1993; Hearst, 1997; Kan et al., 1998; Raynar, 1999). The general notion behind work like this is that there is a connection between the discovery of aboutness or discourse topics and textual organization.

Practical work in topic segment determination goes back to Skorochod'ko

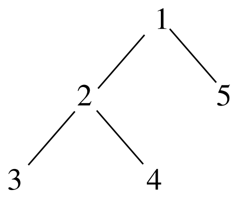
Subjective	*At several different levels, it's a fascinating tale.*
Objective	*Bell Industries Inc. increased its quarterly to 10 cents from seven cents a share.*
Subjective Speech Act	*The South African Broadcasting Corp. said the song "Freedom now" was "undesirable for broadcasting".*
Objective Speech Act	*Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jet-liner's maker are being pursued, a federal judge said.*

Figure 3.26: Wiebe's (1994) Subjectivity Categories

(1972) who makes the connection between topical segmentation and relatedness of terms: whenever the value of “semantic relatedness” of a sentence with respect to the preceding chunk of sentences falls below a threshold, he proclaims a new topical text segment to begin. This idea is taken up in approaches to topic segmentation such as Hearst’s (1997) TextTiling. The assumption is that words which are related to a certain topic will be repeated whenever that topic is mentioned, and that the choice of vocabulary will change when a new topic emerges. Hearst determines boundaries of *topic segments* by calculating vocabulary similarity between two adjacent windows of text. Similarity is defined using the frequency of non-stop word terms in each segment, without taking their inverse document frequency into account. Variations of her approach are discussed in Richmond et al. (1997) where the concepts of *global frequency* and *local burstiness* (proximity of all or some occurrences of multiply occurring content words in a text) are used to refine the definition of segment similarity. Raynar’s (1999) system works by similar principles, but includes a range of other heuristics, similar to the ones used in text extraction methods (cf. section 2.2.1).

Our work is different in its interest in rhetorically, rather than topically, coherent segments. The argumentative zone a sentence belongs to is a distinction which often cuts across subtopic zones. One subtopic might be mentioned in several adjacent argumentative zones. For example, the name of a problem might be repeated in the introduction, in the description of other researchers’ work, the statement which describes weaknesses of that work, in the goal statement and in the description of the own solution. On the other hand, some of our larger zones, particularly the OWN zone, will contain many subtopics. Thus, the apparent similarities between topic segmentation methods and Argumentative Zoning are superficial.

There is a second group of work, providing models of argumentation which have a more general aspiration, analyzing argumentative scientific discourse from a theoretical and logic point of view (Toulmin, 1972; Perelman and Olbrechts-Tyteca, 1969; Horsella and Sindermann, 1992; Sillince, 1992). Argumentation in these approaches is concerned with arbitrary facts about the world and their relation. For a computational treatment to cover this, full text comprehension would be required. Cohen’s (1987) work is more computationally minded. It is a general framework of argumentation for all text types, based on the construction of claim-evidence trees from argumentative text (cf. figure 3.27, taken from Cohen 1987, p. 15):



1	The city is a disaster area
2	The parks are a mess
3	The park benches are broken
4	The grassy areas are parched
5	Returning to city problems, the highways are bad too

Figure 3.27: Cohen's (1987) Evidence-Claim Trees

Argumentative structure in her approach is related to linear order and surface meta-discourse ("clues") like the phrase "*returning to city problems*". Processing is incremental; rules express where in the tree incoming propositions can be attached. This is similar to Polanyi's (1988) discourse grammars where the rightmost node at each level of the tree is always open and all other nodes closed for attachment. Cohen suggests the implementation of a separate clue module within her framework and considers clue interpretation as "not only quite useful but feasible" (p. 18).

Cohen's approach is not implemented. The reason for this is that it presumes a "evidence oracle" which can determine if a certain incoming proposition is evidence for another statement already in the discourse tree. This is a hard task, requiring general inference on the object level which we are trying to avoid at all cost.

An approach for the *generation* of natural language arguments is given by Reed and Long (1998) and Reed (1999). The approach is based on argumentation theory (cf. van Eemeren et al. (1996) for an overview). Their RHETORICA system uses planning to generate persuasive texts by modelling users' goals and beliefs. Apart from the fact that this approach is not concerned with the *analysis* of arguments, the biggest difference between this work and ours is that instead of formally manipulating relations between facts in the world we model *prototypical (fixed) scientific argumentation* in a far more shallow way.

The third group of work related to Argumentative Zoning are discourse theories for rhetorical structure. Discourse structure is concerned with two aspects of the organization of sentences: a) the fact that the sentences in one topical or rhetorical segment of the text are in relation to each other and b) that different segments also have an inter-segmental ordering of intentional relations. This is often referred to as *micro vs. macro-structure* (van Dijk, 1980). Other names for macro-structure are discourse-level structure, or large scale text structure. In a well-written text, the function of micro segments with respect to the macro segment, as well as the function of a macro segment with respect to the text as a whole, is signalled by surface cues. Cues at micro-level are for example connectives between clauses (*"but, thus"*) or enumeration markers (*"first, second, last…"*). Cues at macro level are phrases of the kind *"next we will show that…"*.

We consider here general theories of text structure which are based on intentional or communicative acts of the writer. Examples of rhetorical functions are "to convince a reader", "to provide an example" or "to recapitulate". The common assumption is that in trying to communicate a (set of) messages, e.g., in an argumentative text, humans employ a hierarchical intentional structure.

A bottom-up approach to rhetorical relations, based on a model of human memory organization, is described in the seminal paper by Kintsch and van Dijk (1978). Their main claims about discourse organization are that text content is hierarchical and that relevance is an aspect of discourse organization. Their model starts from a manually-created, logical, but surface-oriented representation for propositions. Connectedness is calculated using the overlap of grammatical arguments in this representation. Even though their theory of text comprehension is plausible, we do not consider it here, as their approach bypasses the essential text analysis phase—this means that it cannot be used for practical summarization of unrestricted text (section 2.2). Instead, we turn to theories which work by considering more superficial cues.

Grosz and Sidner (1986) present a hierarchical discourse structure based on three types of structure: linguistic, intentional and attentional. Intentional structure in their model is defined by those intentions that the writer or speaker intended the hearer to recognize (in contrast to private intentions like to impress somebody). Intentional structure is associated with linguistic units, discourse segments. Two structural relations (dominance and satisfaction-precedence) hold between the segments. In contrast to Swales' model, and similar to Cohen's, an infinite number of different intentions is possible.

Grosz and Sidner state that three kinds of information play a role in the determination of the discourse segments: specific linguistic markers, utterance-level intentions and general knowledge about actions and objects in the domain of discourse. One of their main claims is that the use of certain linguistic expressions like referring expressions is constrained by the attentional structure. The attentional structure contains information about the different possible foci of attention in the conversation: salient objects, properties and relations.

The need to recognize the intentions and their relation to previous intentions is aided in Grosz and Sidner's example, as a strongly hierarchical task-structure underlies their example dialogue. This task-structure provides common knowledge about the task and also acts as a special case of the intentional structure posited.

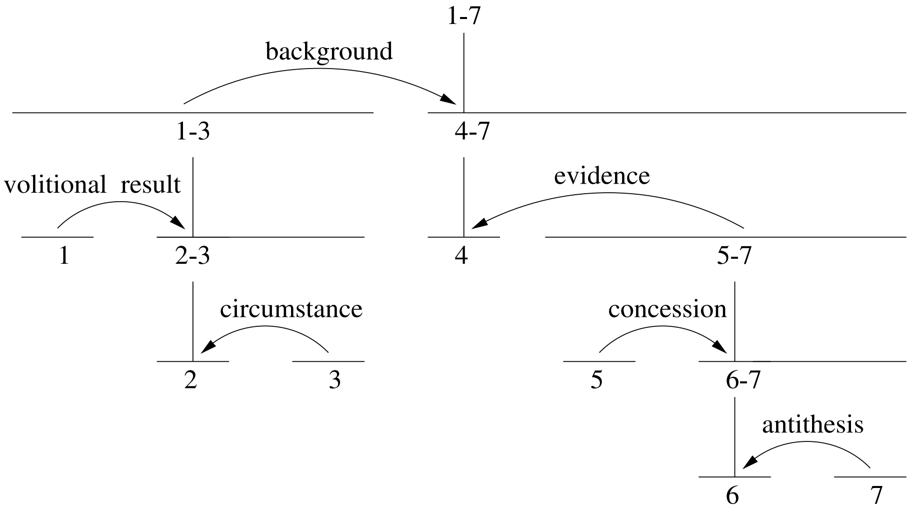
Rhetorical Structure Theory (RST; Mann and Thompson 1987, 1988) is also based on the notion that text structure serves a communicative role. In contrast to Grosz and Sidner, the document structure is based on a *fixed* set of rhetorical relations holding between any two adjacent clauses or larger text segments. Their main claims are that discourse is characterized by strong hierarchical relations and by the predominance of structural patterns of nucleus/satellite type. The relations are typically asymmetric and include CIRCUMSTANCE, SOLUTION-HOOD, ELABORATION, BACKGROUND, ENABLEMENT, MOTIVATION, EVIDENCE, JUSTIFICATION, CAUSE (VOLITIONAL AND NON-VOLITIONAL), RESULT (VOLITIONAL AND NON-VOLITIONAL), PURPOSE, ANTITHESIS, CONCESSION, CONDITION, INTERPRETATION, EVALUATION, RESTATEMENT, SUMMARY, SEQUENCE and CONTRAST. The definitions of the rhetorical relations are kept general on purpose, as illustrated by the one for JUSTIFY:

> JUSTIFY: a JUSTIFY satellite is intended to increase the reader's readiness to accept the writer's right to present the nuclear material.
>
> (Mann and Thompson, 1987, p. 9)

During the analysis, the analyst effectively provides a plausible reason the writer might have had for including each part of the whole text, cf. figure 3.28, taken from (Mann and Thompson, 1987, p. 13–14).

Ambiguity of relations and structure are considered normal in RST (Mann and Thompson, 1987, p. 28). This vagueness poses a problem for computational applications as it leads to multiple RST analyses for a given piece of text. Another dilemma is that researchers building their work on RST have often invented their own, similar relations, such that there was a proliferation of private RST-like schemes; Maier and Hovy (1993) list more than 400 RST-type relations used in the field. This dilemma could be mitigated by a corpus-based approach like Knott's (1996).

Another difficulty is the unit of annotation. It has long been debated, and is still entirely unclear, what the formal linguistic criteria defining such units might be. Consider, for example, unit 7 in figure 3.28 ("*not laziness*"). This unit has been determined



1 Farmington police had to help control traffic today
2 when hundreds of people lined up to be among the first applying for jobs at the yet-to-open Marriott Hotel.
3 The hotel's help-wanted announcement—for 300 openings—was a rare opportunity for many unemployed.
4 The people waiting in line carried a message, a refutation, of claims that the jobless could be employed if only they showed enough moxie.
5 Every rule has its exceptions,
6 but the tragic and too-common tableaux of hundreds of even thousands of people snake-lining up for any task with a paycheck illustrates a lack of jobs,
7 not laziness.

Figure 3.28: Sample RST Analysis

as "clause-like" as it obviously carries a lot of information in this particular argument. However, syntactically, this unit is only a single NP in a VP ellipsis construction—one is now in need of a general syntactic criterion which defines this phrase as a clause, but excludes similar other NPs.

RST has been extensively and successfully used for text generation, e.g. of tutor responses (Moore and Paris, 1993), and of texts describing ship movements and air traffic control procedures (Hovy, 1993). For this purpose Moser and Moore (1996) suggest a synthesis of RST and Grosz and Sidner's theory. On the analysis side, a problem of recognizing RST relations is that most rhetorical relationships are not explicitly marked by connectives, or that it is not clear at which level in the tree a given unit should connect.

Marcu uses heuristics based on punctuation and cue phrases to recognize fully hierarchical RST structure in popular science text (Marcu, 1997a, 1999a,b). One of the applications of the generated structure is summarization. The texts Marcu uses are heavily edited, unlike ours; this makes parsing easier as punctuation can be expected to be standardised. The texts are also well-written: whereas in our texts experts communicate with experts, these texts are aimed at making a possibly non-expert audience understand difficult scientific facts. To do so, causality and other rhetorical relations are often overtly signalled.

Another system that uses RST relations for summarization (of Japanese texts) is BREVIDOC (Miike et al., 1994; Sumita et al., 1992; Ono et al., 1994). Connective expressions in sentences are identified and used to build a representation of the rhetorical relations between sentences. A cumulative penalty scoring technique is used to select the most plausible binary tree. Abstracts of variable length are produced interactively from this structure.

At first glance RST-type rhetorical relations might look a bit like RDP slots, but they have a different status: whereas RST models micro-structure, i.e. relations holding between clauses, RDP slots denote macro structure, i.e. global relations between the given statement and the rhetorical act of the whole article.

While we agree with RST that micro-level structure is likely to be hierarchical and can be well described by RST relations, we choose not to model these relations. For example our move DESCRIBE: OWN SOLUTION, which is particularly long, includes a description of the methodology, evaluation strategy etc. The internal hierarchical structure of this move does not receive any attention in our approach, because we believe that many of the local rhetorical relations between sentences and clauses are irrelevant for our task.

We believe that it is macro-structure and not micro-structure which is useful for summarization and document representation. We also believe that RST is not ideally suited to model macro-structure and that macro-structure is more usefully described by an annotation scheme like ours. When humans are asked to assign RST relations between between paragraphs and larger segments, they often have to resort to the trivial RST relation JOINT. There seem to be fewer constraints on relations between such segments, and we doubt that this structure is hierarchical in the same way that micro-level relations are.

A related fact showing that it is indeed *micro*-level relations that are modelled by RST is the fact that the cue phrases used in RST approaches tend to be connectives, which operate between clauses (Knott, 1996; Marcu, 1997b).

Moreover, even though Mann and Thompson (1987) claim that RST is “unaffected by text size and has been usefully applied to a wide range of text size” (p. 46), RST analysts typically use short texts. Marcu (1997b), for example, uses text with an average of 14.5 sentences, and Mann and Thompson describe a text of 15 utterances as a “larger text” (p. 22)—whereas we wanted to reliably annotate articles several pages long.

To summarize our observations from looking at intention-based accounts, hierarchical intentional relations at micro-level might not be necessary for our task; we believe that global text structure is far more important. Secondly, rhetorical relations between two segments can be recognized by overt clues if they are present. If they are not, there is a problem. The remaining possibilities are the following, all of which are not very appealing:

- One could use simple, short, well-edited texts with standardized punctuation (Marcu, 1997a).
- One could use task-structured texts (Grosz and Sidner, 1986).
- One could posit an “evidence oracle”, i.e., put the task outside one’s remit (Cohen, 1987).
- One could perform “deep” intention modelling and recognition (Pollack, 1986).

In contrast, the task of Argumentative Zoning relies on more superficial expressions of scientific argumentation.

3.6. Conclusion

We have introduced a model of scientific argumentation which describes the argumentative structure of the articles in our corpus. This model incorporates ideas from Swales’ CARS theory of argumentative moves, a certain view on the problem-structure of scientific research and authors’ statements about problem-solving processes, a distinction of contrastive vs. continuative author stance, and our own observations about the attribution of ownership in scientific articles. We have operationalized this model as a 7-pronged annotation scheme. We call the process of applying it to text, i.e. of determining the rhetorical status of each sentence, *Argumentative Zoning*.

We conclude that texts and discourses can have multiple structures at the same time, which are not necessarily isomorphic. Certain structures are particularly dominant in some *text types*, and certain structures are particularly useful for some *tasks*. It seems that for scientific texts our model—relying on fixed, text-type specific argumentative moves—describes one such structure for which both is true at the same time.

The novel aspects of our scheme are that it applies to different kinds of scientific research articles, because it relies on the *form and meaning of argumentative aspects* found in the text type rather than on contents or physical format. It should thus be independent of article length and article discipline.

Other structural descriptions, though useful in their own right, do not fit as nicely to both task and text type: the fixed rhetorical structure of scientific articles, described by models like van Dijk's, Kando's and Kircz', relies on expectations specific to certain domains and therefore cannot describe our data well. General frameworks such as the ones discussed in the previous section, however, do not exploit text type-specific expectations and therefore cannot offer much help for automatic structure recognition.

Figure 3.29 shows the role of RDPs and Argumentative Zoning as intermediaries between reader and writer: whereas RDPs are a representation of what the *reader* wants out of a text (cf. chapter 2), argumentative zones are a representation of what the *author* put into the text.

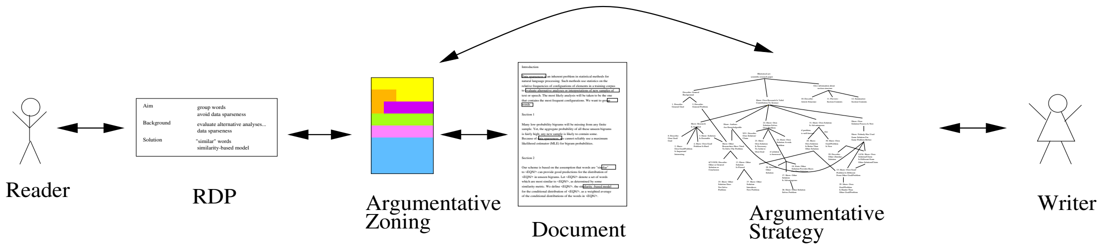


Figure 3.29: Argumentative Zoning and RDPs

This chapter has recast the task of building RDPs as that of Argumentative Zoning. The following questions about Argumentative Zoning now have to be asked:

- How intuitive is Argumentative Zoning? Are the definitions of our categories meaningful to other humans? To answer this question, we observed human annotation with our annotation scheme on naturally occurring, unrestricted text.

We will show in chapter 4 that humans can perform Argumentative Zoning robustly.

- How well can Argumentative Zoning be performed automatically? To answer this question, we built a prototype that applies the scheme automatically, as reported in chapter 5. The results show that Argumentative Zoning can be performed automatically in a robust fashion, although humans are substantially better at the task.

Chapter 4

A Gold Standard for Argumentative Zoning

In the previous chapter, we have introduced a new task: Argumentative Zoning. We will in this chapter define the specifics of the task in such a way that we end up with *gold standards* for it: a definition of what the "right answer" for a set of example documents should look like. For any new task, the right evaluation method is an essential design criterion. Of course, it is essential that the gold standards be defined *before* the experiment, and *independently* of it.

Gold standards are also needed during system development. In chapter 5, we will describe an automatic procedure for determining argumentative zones. We will use our gold standards to determine sentential features and to provide training material. Importantly, gold standards serve for progress evaluation: the evaluation of day-to-day changes to current versions of the system.

Section 4.1 is concerned with finding the right evaluation strategy for Argumentative Zoning. As it is a new task, there is no existing evaluation strategy for it, but the evaluation strategies for similar tasks can inform our decision. We decide to use human judgement; we will then discuss how exactly to define the task in such a way that the similarity of such judgements on the task can be measured objectively.

We will then discuss which numerical evaluation measures to use for the reliability studies (section 4.2). The rest of the chapter is dedicated to describing the reliability studies which measure how much our human annotators agree when they perform Argumentative Zoning.

4.1. Evaluation Strategy

In sections 4.1.1 and 4.1.2, fact extraction and text extraction approaches are contrasted in the light of their evaluation strategies. This contrast will lead to a list of desired properties for our gold standard, and motivate our concrete evaluation strategy for Argumentative Zoning in section 4.1.3.

4.1.1. Evaluation of Fact Extraction

In template-filling tasks like the Message Understanding Conference (MUC; cf. section 2.2.2), the gold standards are called *answer keys*; they are provided by information specialists. Evaluation proceeds by direct comparison of the slot fillers presented by the competing systems with the answer keys.

Answer keys can be of different kinds: they often consist of extracted textual strings, e.g. NPs; sometimes the answer is one of a fixed set of answers ("Was the position newly created, or had it existed before?"). These fixed-choice slots often require inference from subtle linguistic cues. Slots can also be filled by pointers to other templates, or may contain numerical values which the systems have to calculate if the values are not present in the text.

It is easy for humans to assess the correctness of these answer keys after having read the text, as the slot semantics is concrete and domain-specific. However, even though humans can *decide* whether an answer key is correct or not, it is still not an easy task for human experts to *fill* template slots consistently. For more complex slots, there might be two different, but equally appropriate ("correct") keys—superficially different material, coming from different places in the document.

Sometimes, there is an overlap problem, e.g. when one annotator decides to include an apposition of an NP in a slot and the other does not. Annotation guidelines (Chinchor and Marsh, 1998) provide decision criteria for this and other problematic cases.

To measure how often annotators disagree, a subset of the materials (about 30% of the texts), is provided with answer keys by more than one expert. The keys of one annotator are taken as gold standard in turn, and percentage agreement is calculated, i.e. the percentage of identical keys over total keys. A full discussion of evaluation measures for tasks like this is given in section 4.2. In MUC, only reproducibility is reported (e.g. 83% for Scenario Templates); no stability tests are conducted, i.e., it is not measured if the same annotator will annotate in a similar way at a different point in time.

There are disadvantages associated with this type of gold standard. A simple comparison of *one fixed* answer key does not incorporate enough flexibility to deal with cases where the system's answer is different from the answer key. As we cannot perform deep understanding, we need a fair comparison method which deals with *surface* strings. Direct surface comparisons might punish the system unfairly: the answer might be a string which looks different but means something very similar to the given answer key. Fairer system evaluation should give the system a score better than zero in case a second-best answer is retrieved by the system instead of the best answer. What is needed is a gold standard which can provide some kind of fall-back option, i.e. other acceptable—albeit less relevant—answers.

4.1.2. Evaluation of Text Extraction

Gold standards consisting of whole sentences—*target extracts*, i.e. a set of sentences that together constitute the best possible extract from a document—are still the most typical gold standard for text extraction-type summarizers, as target extracts allow for a simple comparison with the machine produced extracts. The problem of evaluation seems to get simpler when gold standards are always full sentences; at least, there is no overlap problem, as there might be with MUC answer keys.

There are different methods whereby one could achieve a target extract, e.g. by asking humans to select important sentences from the text, or by finding other independent, objective criteria for "extract-worthiness", e.g. similarity of document sentences with sentences in a human-written abstract.

4.1.2.1. Free-selecting Sentences from Documents

Early researchers developing corpus resources for summarization work have often defined their own target extracts, relying only on their intuitions (see, e.g. (Luhn, 1958; Edmundson, 1969)). Some have tried a more objective approach by asking unrelated humans to prepare a target extract, i.e. subjects which are not involved in the process of automatic summarization. Several researchers report reasonable agreement between their subjects (Klavans et al., 1998; Zechner, 1995) for free-selecting sentences from newspaper text.

Using unrelated subjects, however, still does not guarantee objectivity: Paice and Jones (1993) reject the use of free-selected sentences for the evaluation of their template-generated summaries, as a small trial showed that their (expert) subjects' selection strategies were very heavily biased towards their individual research interests.

The texts chosen are typically short, so that there are few alternative sentences that *could* have been chosen by the subjects, and the journalistic style makes the selection easier still: the most important sentences will be found in the beginning (Brandow et al., 1995).

For scientific text, the level of subjectivity needed for the task might be higher. Rath et al. (1961) report low agreement between human judges carrying out free selection. If six subjects were asked to select 20 sentences out of *Scientific American* texts ranging from 78 to 171 sentences, all six of them agreed only on 8%, and five agreed on 32% of the sentences. Rath et al. also found that annotators only chose 55% of the sentences they chose six weeks ago. Edmundson (1961) reports similarly low human consistency.

The text extraction evaluation strategy also suffers from surface comparability problems: an ideal gold standard should treat two or more sentences in the text alike, if they express the same semantics. However, target extracts do not account for the cases where two sentences are directly replaceable, or where two sentences taken together contain roughly the same information as another one. There is not a single best target extract for a document:

> [the] lack of inter- and intra subject reliability seems to imply that a single set of representative sentences does not exist for an article. It may be that there are many equally representative sets of sentences which exist for any given article. (Rath et al., 1961, p. 141)

4.1.2.2. Abstracts as Gold Standards

One would ideally want a gold standard which allows different research teams to replicate the gold standard. Asking humans to select sentences does not provide this level of objectivity, of course, as relevance is situational (cf. section 2.1). Researchers have thus looked for an independent, fixed definition of relevance which comes with the text itself and which cannot be influenced anymore, e.g. one that is based on a historic decision of a professional (the indexer or the abstractor). Such a gold standard could be given by a back-of-the-book index (Earl, 1970), or by the human-written abstract (Kupiec et al., 1995).

Earl used a back-of-a-book index to identify all sentences in a book chapter that contained an indexed term; these *indexible* sentences constitute her gold standard. But scientific articles do not typically contain back-of-a-book indexes. Kupiec et al. (1995) use the summary supplied with the article instead to define the gold standard sentences: their gold standard is the set of sentences in the source text that are maximally similar ("align") with a sentence in the summary. An automatic similarity finder is used to identify potential pairs of summary and source text sentences by superficial criteria; subsequently, a human judge (presumably one of the system developers) decides if the alignment is justified on semantic grounds. For alignment to hold, Kupiec et al. allow for minor modifications between sentences; full matches, partial matches and non-matches were possible.

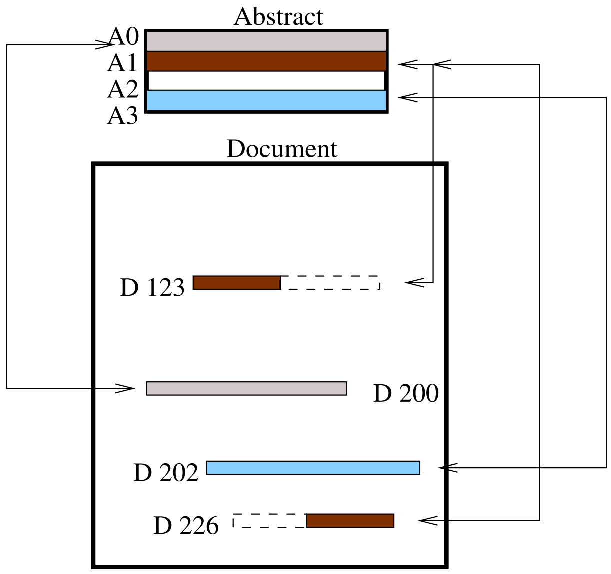


Figure 4.1: Target Extract by Alignment (Kupiec et al., 1995)

In Kupiec et al.'s corpus of 188 engineering articles plus summaries, 79% of the sentences in the summary could be aligned with sentences in the source text. In figure 4.1, for example, document sentences D-200 and D-202 align with abstract sentences A-0 and A-3, respectively. Parts of sentences D-123 and D-226 align with abstract sentence A-1, whereas abstract sentence A-2 does not have a corresponding sentence in the document. Examples for matches and non-matches from our corpus follow; they were obtained in a duplication of Kupiec et al.'s experiment (cf. section 5.3.4.1; also described in Teufel and Moens 1997).

> *Summary:* In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent. (A-3, 9405013)

> *Document:* An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent.
> (S-131, 9405013)

The previous sentence pair illustrated a *match*, the following sentence pair a *non-match*:

> *Summary:* Recent studies in computational linguistics proposed computationally feasible methods for measuring word distance. (S-2, 9601007)

> *Document:* The paper proposes a computationally feasible method for measuring context-sensitive semantic distance between words. (A-0, 9601007)

The last example illustrates one of the rare cases where syntactic similarity does not mirror semantic similarity: however similar, the sentences have different propositional content, as one refers to previous work and the other to the work discussed in the source text itself.

Gold standard definition by abstract similarity is attractive because the machinery is technically simple, and the definition solves the objectivity dilemma: gold standards are defined by an independent method which is in principle outside the system developers' control. Correcting the automatically determined alignment—the only point where the system developers interact with the gold standards—requires relatively little human intervention and introduces little subjectivity. Kupiec et al. even argue that gold standards attained by *uncorrected* alignment are almost as good for system training as the corrected ones. Subsequently, the idea of using the abstract as gold standard has found a number of followers (Mani and Bloedorn, 1998; Hovy and Liu, 1998).

However, Kupiec et al.'s method introduces a dependency on the quality of the abstracts, and on the process of how they were generated. This is an issue with our texts. In our corpus, the abstracts were not written by professional abstractors but by the authors themselves.

While the literature on summarization techniques for professional abstractors is large (cf. section 2.1.1 and 2.3.1.2), there is not much research into how non-information specialists generate abstracts. However, it is indeed commonly assumed that author summaries are of a lower quality when compared to summaries by professional abstractors (Lancaster, 1998; Cremmins, 1996; Rowley, 1982). Rowley says about author abstracts that they are sometimes poorly written, that they often contain too much or too little data, and that there is often undue emphasis on author's priorities.

Borko and Bernier (1975) similarly caution that authors do not necessarily write the best abstracts for their papers, and Dillon et al. (1989) found empirically that journal-scanning readers often ignore author-written summaries if the full article is available too, and reject the summaries as "misleading" or "biased". We see several dangers with author summaries as gold standards for our task:

- We suspected that there is a less systematic relation between the information contained in the author-written summaries and the information contained in the documents. If it is not the case that the abstracts were created predominantly by selecting sentences, but if they were created from scratch, a surface-alignment procedure might provide too few gold standard sentences, and coverage would be too low, and indeed, this is the case in our corpus. Authors tend to reuse less of the document sentences, but *deep generate* new sentences from scratch.

- The papers in our collection come from different presentation styles, academic traditions and cover a wide range of subdomains. As a result, they differ in their internal document structure.

- They also differ in the structure of their abstracts. There is no guarantee that abstracts written by the authors keep to any kind of fixed rhetorical building plan, which abstracts produced by professional abstractors do (Liddy, 1991). Even though the information which ends up in the author abstracts is most certainly *relevant*, there are large individual differences of style and preference with respect to what *kind* of information an abstract contains, particularly if the authors of the abstracts were careless or biased. In a task such as ours it is essential that if there is information which is of comparable rhetorical status across papers, then the gold standard should mark this information similarly, independently of presentation form or where in the paper the information occurs. Comparability of information is hard to obtain with a surface-based method anyway, but if author decisions are taken to define the gold standard, comparability across papers decreases dramatically.

  Indeed, a later analysis (cf. section 4.4.1) reconfirms that the length and structure of our author abstracts vary considerably from paper to paper.

- Abstracts written by professional abstractors are typically self-contained, such that they can be understood without reference to the full paper. In many examples in our materials, this is not the case.

- Even worse, it is not even guaranteed that all the information contained in the abstract will also occur in the main document in *some* form. Writing advice states that the text and the abstract, apart from conveying the same semantics, should be viable texts which can be read on their own. But some of the authors in our collection assumed that the abstract would always be read before the main document, and in order to save time, they “abused” the abstract as an introduction. We found five papers in our collection where information in the abstract is not repeated anywhere else in the main document. Such cases are catastrophic for approaches which derive their gold standard from the abstract.

In early experiments with alignment (Teufel and Moens, 1997), we use a simple surface similarity measure which computes the longest common subsequence (LCS) of non-stop-list words. The results show a much lower alignment rate of 31% in our corpus, in comparison to Kupiec’s 79%.

For example, consider the author summary of our example paper and the best-aligned sentences (figure 4.2).

Sentence **A-2** does not align with any document sentence, and alignments **A-1–113** and **A-3–147** were rejected by the human judge (us) as bad matches. The one acceptably aligned abstract sentence (**A-0**) is only partially aligned—with sentences **0** and **164**. Overall, the authors do not seem to have prepared the abstract by sentence extraction: all abstract sentences are at a higher level abstraction level than the corresponding document sentences, cf. the difference between **A-3** and **147**. It is immediately clear from the low level of alignment that this particular target extract cannot be a good representation of the document, even though the author abstract itself is.

Matters get even more complicated when we look at the rhetorical status of sentences, which is essential for Argumentative Zoning. For example, the rhetorical structure of the original abstract consisted of a sequence of Research goal (**A-0**), Solution applied (not invented by Pereira et al.; **A-1**), Further description of the solution (**A-2**), and Description of the evaluation (**A-3**). This summary is most similar in type to the summary for intellectual ancestry for uninformed readers, as discussed in section 2.3.3 (figure 2.20, p. 69). In comparison to the original abstract, the target extract is impoverished with respect to rhetorical structure; it consists of a very general statement about the task, and a statement that a solution was found—only 2 out of the 7 slot fillers available in the author abstract. Even though the aligned document sentences might be superficially similar to the ab-

	Abstract sentences		Aligned document sentences
A-0	We describe and experimentally evaluate a method for automatically clustering words according to their distribution in particular syntactic contexts.	**0**	`(partial)` Methods for automatically classifying words according to their contexts of use have both scientific and practical interest.
		164	`(partial)` We have demonstrated that a general divisive clustering procedure for probability distributions can be used to group words according to their participation in particular grammatical relations with other words.
A-1	Deterministic annealing is used to find lowest distortion sets of clusters.	**113**	`(bad match)` The analogy with statistical mechanics suggests a deterministic annealing procedure for clustering <REF>Rose et al. 1990</REF>, in which the number of clusters is determined through a sequence of phase transitions by continuously increasing the parameter <EQN/> following an annealing schedule.
A-2	As the annealing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data.		—
A-3	Clusters are used as the basis for class models of word coocurrence, and the models evaluated with respect to held-out test data.	**147**	`(bad match)` For each critical value of <EQN/>, we show the relative entropy with respect to the asymmetric model based on <EQN/> of the training set (set train), of randomly selected held-out test set (set test), and of held-out data for a further 1000 nouns that were not clustered (set new).

Figure 4.2: Author Abstract and Target Extract by Alignment for Document 9408011

stract sentences, their rhetorical status is not necessarily similar to that of their aligned sentences. Without context, the rhetorical status of the document sentences cannot be detected anymore, and it is not even clear that it would be of help. Clearly, something got lost on the way.

To sum up, we do not deny that there are cases where abstract alignment can define good gold standards, and that Kupiec et al.'s experiment is probably one such case. However, the reason for this is not that abstracts *per se* provide good definitions of gold standards—rather, it is due to other fortunate circumstances, like the extensive training of professional abstractors and the high homogeneity with respect to paper and abstract structure in some data collections. In our case, alignment with abstracts would probably define a low-quality gold standard.

In general, surface comparability remains a problem for target extracts by abstract similarity. Document sentences which share propositional contents with an abstract sentence but which look different on the surface will not be contained in the gold standard, even though they should be; and a system which correctly determines such a sentence would be unduly punished by a target extract gold standard.

There is an additional problem with the static nature of the gold standard definition. The fact that the gold standard cannot be touched anymore might well make it more "objective", but the process by which the abstract was obtained (described in abstractors' guidelines) does not necessarily provide the specific information needed for a given task. For example, our task demands finding information about the goal of the paper, in relation to previous work: the determination of rival approaches and supporting previous research is essential. Unfortunately, this type information is not traditionally present in abstracts. Instead, this information might be hidden *anywhere* in the texts. An advantage of asking subjects to free-select sentences is that new criteria can be applied to the search as needed, in a *dynamic* way. As these criteria of selection are defined after the creation of the text, they can be changed according to task requirements.

One good point about target extracts in general, no matter by which method they are obtained, is that only a small number of sentences are selected, which are guaranteed to be globally important. This would be an advantage for a gold standard like ours. But most target extracts do not provide fall-back options, as they only make a binary distinction between relevant and non-relevant sentences.

4.1.3. Our Evaluation Strategy for Argumentative Zoning

We have argued that neither target extracts nor MUC-style answer keys can offer us high-quality gold standards for our task. But there is yet another field whose gold standards might be important for us.

Gold standards by total-coverage are traditionally in use in areas where the annotation in the text serves as a long-term resource itself, e.g. in dialogue act coding

(Carletta et al., 1997; Alexandersson et al., 1995; Jurafsky et al., 1997). Humans are asked to classify utterances in a corpus into a finite set of categories (called "dialogue moves"). For this kind of exercise, it is essential that different annotators performing the task independently (e.g. in different places) can create a resource that fits in with already existing resources generated according to the same annotation scheme. High reproducibility of such a scheme is thus important. Training of the annotators is extensive. Guidelines, also referred to as *coding handbooks*, are used to describe the task and the semantics of the categories. Specialized, standardized statistics, borrowed from the content analysis community (Carletta, 1996), exist for testing certain properties of the annotation scheme, most notably stability and reproducibility (cf. section 4.1.1).

Our evaluation strategy is as follows: we elicit judgements from subjects about the argumentative status of sentences in the source text according to our annotation scheme. Subjects perform *full-coverage* annotation, i.e., they give a judgement for *each* sentence in the paper.

We argue that this evaluation strategy will improve the gold standard situation with respect to surface comparability, fall-back options and comparability between papers. System evaluation is no longer a comparison of extracted sentences against a finite set of "good" sentences—this inevitably cannot work because there is not just *one* possible extract for a paper. Instead, *every* sentence in the source text which expresses the main goal will have been identified, and the system's performance is evaluated against *that* classification, providing an evaluation that portrays the real situation better.

We have, in chapter 3, made an implicit claim about the adequacy of our annotation scheme: that its categories provide an intuitive description of certain aspects of scientific texts. But the semantics of our slots are not as simple as the domain-specific MUC slots, which have the advantage that humans can confirm with high confidence if a slot filler is "right". In our scheme more subjective judgements are necessary. If we could prove a high degree of human agreement on the application of argumentative zones, this would also serve to verify our definition of the zones. Learnability of the scheme (and, as a result, reasonable reproducibility) is also important from a practical point of view as we want to use the gold standards as training material if they constitute a reliable resource.

The main difference between our task and other total-coverage annotations is that our task is a document retrieval task, and as a result, relevance *is* an issue for us. Certain items are more important for us than others, and certain errors are more grave than others. We care most about reproducibility in those zones which are particularly important for our task (e.g. AIM zones); we care less about errors in the frequent zones as these sentences are not directly extracted and displayed in RDPs.

Our gold standard should give us sentences which are the *best* slot fillers for each category; it should also define fall-back options. However, total-coverage classification does not readily provide different degrees of relevance. It gives us many "equally relevant" sentences per category, whereas the other gold standards would have given us few "relevant" sentences. In an independent step, the most appropriate slot fillers would have to be determined:

1. Subjects could tell us which sentences are the best fillers, e.g. by ranking their prior classifications.

2. Some external criterion could define relevance independently of the human classification; e.g. sentences alignable with abstract sentences or occurring in the periphery of the paper could be considered "more relevant". The connection between location and the quality of gold standards is explored in section 4.4.2.

3. Slot fillers which are similar to each other could be defined to be more relevant. This approach was suggested in section 2.3.3 where we sketched the generation of tailored summaries from RDP slot fillers.

Apart from not being able to give us the most appropriate slot fillers, total-coverage classification gold standards provide a well-suited evaluation for our task, as such classification is a simple, well-understood cognitive task with a widely accepted evaluation metrics. However, it is a time-consuming task—we consider different ways of reducing the effort, either by reducing the training (cf. section 4.3.2) or by reducing the areas to be annotated (cf. section 4.4.2). Our new gold standard helps us get around some of the problems that other evaluation strategies have:

- *Objectivity:* The new gold standard measures objectivity in terms of stability and reproducibility, i.e. in how far humans will agree on the task (results are reported in section 4.3). One could, however, argue that a static, fixed, independent standard as in Kupiec et al.'s work is intrinsically more objective.

- *Task-flexibility:* Instructions to the annotators can be adjusted according to the requirements of the task.

- *Comparability between papers*: The new gold standard guarantees comparability because *all* sentences are classified. Coverage of all categories should be high, i.e. there should always be enough candidates for each category. As a result, a sensible comparison of information between papers is possible, unlike in the abstract-as-gold-standard strategy.

- *Fall-back options:* The new gold standard provides fall-back options for each category (provided the category was present in the paper), unlike other methods.

- *Best fillers:* The new gold standard still gives too many fillers per category, all of which are judged equally-relevant, in contrast to selection methods. In order to determine the most relevant fillers in our case, an independent measure of relevance is needed.

- *Surface comparability:* The new gold standard has fewer problems with surface comparability than target extracts or answer keys. This is due to the fact that judgements for each sentence are compared.

4.2. Evaluation Measures

In the following experiments, we are particularly interested in two properties of our annotation scheme: Firstly, stability, i.e. the extent to which one annotator will produce the same classifications at different times (Krippendorff, 1980). Stability is important, because in unstable annotation schemes the definition of the categories is not even consistent within one annotator's private understandings, and as a result, such schemes are very unreliable. High stability shows at the very least that there must be *some* consistent definition of semantics in the gold standard, even if we do not know yet if this definition can be communicated to others. The second property is reproducibility, i.e. the extent to which different annotators will produce the same classifications, which measures the consistency of shared understandings (or meaning) held by more than one annotator. As consistent *shared* understandings require consistent *private* understandings, an unstable annotation can never be reproducible; conversely, it is commonly assumed that a proof of the reproducibility of a scheme implies its stability. Thus, many experimentators only measure and report reproducibility (cf. the MUC enterprise, section 2.2.2).

We feel that stability is independently important, and that stability and reproducibility have completely different consequences with respect to our task. Researchers in document retrieval have argued that although stability is important to some degree, if one is interested in user satisfaction, then reproducibility is of little importance. If there are two or more intuitively "good" but different gold standards, two judges might disagree over which one to choose, resulting in a low reproducibility. However, both of these gold standards might have satisfied the user. We subscribe to the argument of theoretical priority of stability over reproducibility in document retrieval, but at the end of the day, only extrinsic evaluation can prove or disprove if the argument is valid.

A related question is how exactly we should establish an *upper bound* for the task. An upper bound is the best measurement that an automatic performance can *theoretically* reach. When humans systematically do not agree beyond a certain degree, this degree must be accepted as the upper bound: it makes no sense to think of a machine as performing better than this level of agreement. We argue that reproducibility constitutes a good upper bound. That is, if the performance stays the same if an automatic approach is added to a pool of independently annotating human annotators, then this approach has reached the theoretical best performance possible.

In many related tasks, definitions of upper bounds are handled less strictly. Kilgarriff (1999), for example, reports an upper bound for word sense disambiguation which is numerically very high. This gold standard was gained by negotiation between the annotators, as is common in lexicography. We also believe that interaction between annotators is important, in order to arrive at a shared understanding of the categories. However, experience has shown that it is often the annotator with the strongest personality which convinces the other annotators of the validity of her annotation.

Another form of improving "reproducibility" would be to ask annotators to *correct* somebody else's output—in other tasks like manual parts-of-speech (POS) assignment, annotators have been shown to agree much more if they do not perform the task from scratch.

However, as we are interested in the properties of the cognitive task, we measure reliability of independent annotation *before* discussions. The real keepers of the semantics of the categories should always be the guidelines. The guidelines for annotation tasks should be written before the experiment and changed as little as possible during the experiment. However, as annotation experiments are long and expensive enterprises, it might be difficult to repeat an experiment after each change (and ideally with new annotators). We had to change the guidelines several times (e.g., the example annotations in figures on p. 327 and 328 were added after those papers had been annotated independently).

Our annotation task is mutually exclusive categorial assignment. There have been different ways in the past to evaluate agreement between humans for such task (cf. the overview in Carletta 1996), using either majority opinion or percentage agreement as measurement. We are opposed to using majority opinion: the average does not reflect anybody's understanding of the categories. We want to treat all our annotator's opinions as a valid judgement. None of these is by definition wrong or right—we are dealing with a difficult "high-level" task, where a certain level of subjective disagreements can be expected.

We use the Kappa coefficient K (Siegel and Castellan, 1988) to measure stability and reproducibility among k annotators on N items (here: sentences). For our task, Kappa has the following advantages:

- It factors out random agreement.
- It allows for comparisons between arbitrary numbers of annotators and items.
- It treats less frequent categories as more important.

The Kappa coefficient controls agreement $P(A)$ for agreement by chance $P(E)$:

$$K = \frac{P(A)-P(E)}{1-P(E)}$$

No matter how many items or annotators, or how the categories are distributed, $K = 0$ when there is no agreement other than what would be expected by chance, and $K = 1$ when agreement is perfect. If two annotators agree *less* than expected by chance, Kappa can also be negative. Chance agreement is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators. Kappa is stricter than percentage agreement: its value is always lower or equal to percentage agreement P(A); it is equal in the case of a uniform distribution and lower for skewed distributions. We already know that our category distribution will most likely be very skewed, for example because the category OWN is so predominant. The fact that Kappa is a more sensible measurement for our task than percentage overlap can be easily shown with the following argument about baselines for our task. (This argument anticipates some numerical values which we will obtain later on in this chapter.)

Choosing the most frequent category OWN is one possible baseline for our task (Baseline 1). Figure 4.3 shows that percentage agreement makes this baseline look like a good one at 69%, in comparison to human agreement at only 87%. However, if Kappa is used to measure the similarity of this baseline with the annotation of a human annotator, it reveals a negative (K=–.12)—compared to chance agreement, the baseline performs *worse* than random. This agrees with our intuition that always choosing the most frequent category is a bad strategy for our task. For our task it is important to choose the rare categories AIM, TEXTUAL, CONTRAST and BASIS.

Baseline	Kappa	P(A)	P(E)
Baseline 1: Most frequent category	-.12	68%	71%
Baseline 2: Random, uniform distribution	-.10	14%	22%
Baseline 3: Random, observed distribution	0	48%	48%

Figure 4.3: Baselines for the Task of Argumentative Zoning

We implemented a random generator, assigning categories based either on a uniform distribution (Baseline 2) or the observed distribution (Baseline 3). Baseline 2 has a slightly better chance agreement; it achieves K=–.10 if compared to the human annotator. The hardest-to-beat baseline is random choice according to the observed distribution of categories (Baseline 3). Kappa for this baseline should theoretically be K=0 which is reconfirmed by our data. Kappa agrees with our intuition that Baseline 3 is better than Baseline 1 whereas the numerical values of percentage agreement contradict our intuition.

Kappa is designed to abstract over the number of annotators as its formula relies on *pairwise* agreement. That is, K for $k = 6$ annotators will be an average of the values of K for $k = m$ where $m < k$, taking all possible m-tuples of annotators from the annotator pool. This property makes it possible to compare between different numbers of annotators, and between groups of annotators and versions of our system. A look at Rath et al.'s awkward way of reporting agreement for different annotator pools (cf. p. 132) makes clear that numerical comparability is a big advantage.

We are also looking for a measurement which will punish disagreement on the rare (= important) categories more than disagreement in the more frequent categories. As a side effect of taking random agreement into account, Kappa treats agreement in a rare category as more surprising, and rewards such agreement more than an agreement in a frequent category.

There are different scales of how to interpret Kappa values. Krippendorff (1980) starts from the assumption that there are *two* independently annotated variables which show a clear correlation. If the agreement of an annotation of one of these is so high that it reaches a value of K=.8 or above on a reasonably-sized dataset, then the correlation between these two variables can be shown with a statistical significance of $p \leq 0.05$. That is, the annotation contains enough signal to be found among the noise of disagreement. If agreement is in a range of $.67 \leq K < .8$, the correlation can be shown with a (marginal) statistical significance of p=0.06, which allows for tentative conclusions to be drawn. Krippendorff's strict scale considers annotations with $K < .67$ as unreliable. More forgiving scales take into account that most practical annotation schemes only mark one dependent variable and assume that K=.6 is still reasonable agreement. However, Krippendorff (1980, p. 147) describes an annotation experiment performed by Brouwer et al. (1969) in which annotators achieved K=.44 with an annotation scheme whose categories were described only by complicated Dutch names with no resemblance to English words. This is disturbing, because Kappa *should* have been zero, due to the lack of semantics attached to the categories (as the annotators did not understand Dutch): any agreement achieved in that experiment can be only considered as chance. Having said this, it is so difficult to achieve *high* Kappa values that one can nevertheless exclude chance in those cases—Kappa is in general accepted in the field as a sensible and rigorous measure.

Whereas researchers using Kappa frequently have developed some intuitions about whether or not not two Kappa values probably are statistically significantly different or not, there still is no statistical formula to calculate if this is the case or not. This is a disadvantage of using Kappa, but we think it is out-weighed by its advantages.

We use our own implementation of Kappa which allows us to vary annotation areas (cf. section 4.4.2), calculate values for single files, subsets of annotators in the pool and to show confusion matrices for pairs of annotators.

4.3. Reliability Studies

4.3.1. Experimental Design

We conducted three studies. The first two, studies I and II, were designed to find out if two versions of our annotation scheme can be learned by human annotators with a significant amount of training. The first version is the *basic* annotation scheme which encodes intellectual ownership (cf. section 3.3). The second version is the *full* annotation scheme with seven (more complicated) categories. A positive outcome of studies I and II would convince us that the human-annotated training material constitutes a good gold standard, and that it can be used for both training and evaluation of our automatic method in chapter 5. The outcome of study II is crucial to the task, as it deals with the full annotation scheme. Some of the categories specific to the full annotation scheme (AIM, TEXTUAL, BASIS and CONTRAST) provide essential information for RDPs.

Study III tries to answer the question if the considerable training effort used in studies I and II can be reduced. If this were the case, i.e. if annotators with no significant task-specific training could produce similar results to highly trained annotators, the training material could be acquired in a more cost and time effective way. A positive outcome of study III would also substantiate claims about the immediate intuitivity of the category definitions.

4.3.2. Study I

4.3.2.1. Method

Subjects: Three annotators participated in this study: Annotator A holds a Master degree in Cognitive Science and Annotator B was a student of Speech Therapy at Queen Margaret's College, Edinburgh. Annotator C is the author of this thesis. The annotators can be considered skilled at extracting information from scientific papers but they are not experts in all of the subdomains of the papers they annotated. Annotator A has some overview knowledge in most of the subfields represented in the corpus; in particular, he is well accustomed to articles in computer science, which Annotator B was not. Annotator B had some knowledge in phonology and phonetics, and to a lesser degree in theoretical linguistics. Annotators A and B were paid for their work at the standard academic student rate of the University of Edinburgh.

Materials: The materials consist of 26 computational linguistics papers from our collection (cf. appendix A.2 for the overall list of articles in our corpus). Figure 4.4 lists the materials used in this study: the papers and their numbers of sentences (abstract sentences and document sentences, but excluding sentences occurring under the heading *Acknowledgements*). We used the first four articles of our collection (papers 0 – 3) for training, and the next 22 papers (papers 4 – 25) for annotation by all three annotators. As we wanted to cover as much variety as possible in writing style, we decided to only include one paper by each first author in each study—subsequent papers by the same authors were discarded. In study I, no paper was excluded on the grounds of authorship, however. During the annotation phase, one of the papers (paper 18) turned out to be a review paper. This paper caused the annotators difficulty as the scheme was not intended to cover reviews. Thus, we discarded this paper from the analysis. For the stability figures (intra-annotator agreement), 5 papers were randomly chosen out of the set of 21 papers.

Type of Material	Paper numbers	Sent.
Training material	0, 1, 2, 3	532
Annotation material	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25	3643
Intra-annotator material	0, 7, 10, 23, 24	1115

Figure 4.4: Study I: Materials

Procedure: The training procedure was as follows: the annotators read our written instructions which define the categories of the *basic* version of the annotation scheme in detail (7 pages; reproduced in appendix C.1). For the reader's convenience, figure 4.5 repeats the categories of the basic annotation scheme.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...

Figure 4.5: Study I: Overview of Basic Annotation Scheme

After reading the guidelines, the annotators marked up the first two training papers, followed by a discussion, then the other two training papers, followed by another discussion. In these discussions, we tried to settle disagreements in the annotators' judgements and change unclear passages in the instructions.

The annotation procedure itself was as follows: Annotators marked up the 21 papers, 5–6 papers per week, in the same order. There was no communication between

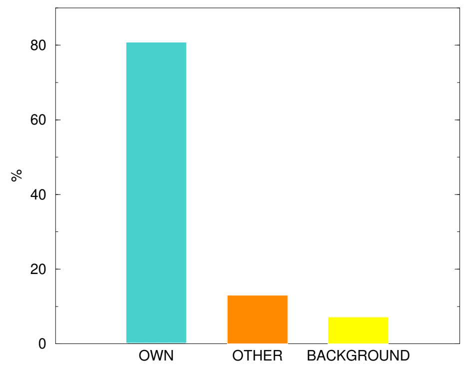


Figure 4.6: Study I: Overall Frequency of Categories

the annotators during the annotation. Annotation included the abstracts as well as all sentences in the document (excluding acknowledgement sentences). Reading and annotating a paper took the annotators 20–30 minutes on average. Weekly discussions between the three annotators took place during the annotation phase. The rationale of the discussions was to increase *future* agreement by clarify unclear passages in the guidelines in the light of unclear annotation cases. However, agreement was measured *before* discussions. As there was no time to implement a specific annotation tool, all annotation reported here was done pencil-on-paper and then edited into an XML version of the documents.

6 weeks after the end of the first annotation phase, stability was measured by an intra-annotator experiment, where annotators were asked to re-annotate randomly chosen papers.

We collected informal comments from our annotators about how natural the task felt, but did not conduct a formal evaluation of subjective perception of the difficulty of the task. Instead, our analysis concentrates on trends in the data as the main information source.

4.3.2.2. Results and Discussion

The results show that the basic annotation scheme is stable (K=.83, .79, .81; N=1115; k=2 for all three annotators) and reproducible (K=.78, N=3643, k=3). This reconfirms that trained annotators are capable of making the basic distinction between own work, specific other work, and general background. To our knowledge, this study is the first to research attribution of intellectual ownership empirically on a corpus.

Categories			Kappa
OWN + OTHER 93.2%		BACKGROUND 6.8%	.58
OWN 80.4 %	OTHER + BACKGROUND 19.6%		.83
OWN + BACKGROUND 87.2%		OTHER 12.8%	.77

Figure 4.7: Study I: Krippendorff's Diagnostics for Category Distinction

Figure 4.6 shows that the distribution is very skewed, as predicted. The relative frequency of the three categories is 80.4% (OWN), 12.8% (OTHER) and 6.8% (BACKGROUND).

Though the reliability values are acceptable, there are some questions that are typically asked in order to improve an annotation scheme:

- Do all annotators perform equally well?
- Are there particular category distinctions that are hard to make?
- Is there a difference between clusters of items (papers)?

The first question is answered easily—the variation between annotators is fairly small. The results for pairwise comparison are K=.74 (A, B), K=.78 (B, C) and K=.82 (A, C). It is important that the results do not change dramatically when the developer of the annotation scheme (Annotator C) is left out of the annotator pool. In this case, they drop a little from K=.78 to .74. This still suggests that the training conveyed the intentions of the developer of the annotation scheme fairly well.

In order to see which category distinctions are hard to make, we use Krippendorff's diagnostic for category distinctions: all other categories but the one(s) of interest are collapsed. The most difficult single distinction is the one that results in the *best* reproducibility values if omitted. In our case, this most difficult distinction is the one between OTHER and BACKGROUND. We are not surprised about this: the distinction between other general work and other specific work concerns only the degree of specificity. Swales (1990) reports similar difficulties with a distinction between his two related moves 1.2 (making topic generalizations; background knowledge) and 1.3 (reviewing previous research). There might not be an easy way to avoid this difficulty; it seems to be part and parcel of the task.

Figure 4.8 shows that the variation in reproducibility across items (papers) is large: there are some papers that are annotated very consistently, and others that are not.

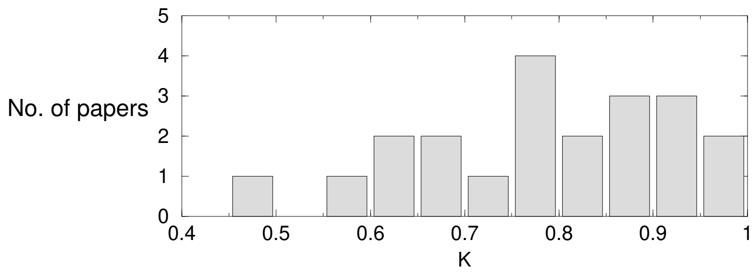


Figure 4.8: Study I: Distribution of Reproducibility Values

We tried to diagnose the reasons for the low reproducibility of some papers. We have several hypotheses of what could be responsible for this:

1. One frequent problem our annotators reported was a difficulty in distinguishing OTHER work from OWN work, due to the fact that some authors did not express a clear distinction between *previous* own work (which, according to our instructions, had to be annotated as OTHER) and *current, new* work. Our annotators reported that in some papers there are long sections that cannot be obviously attributed to either *previous* or *current* work because the authors did not make the distinction clear. This was particularly the case where authors had published several papers about different aspects of one piece of research (cf. the idea of "smallest publishable unit", section 3.2.3).

   We suspected that the effect of mixing descriptions of own and previous research could be gauged by the *self citation ratio*, i.e. the ratio of self citations to all citations in running text. 5 papers contain no self citations and were thus put into one group. We divided the remaining papers into two equally sized groups, one with a high and one with a low self citation ratio (the borderline turned out to be at 18% of all citations).

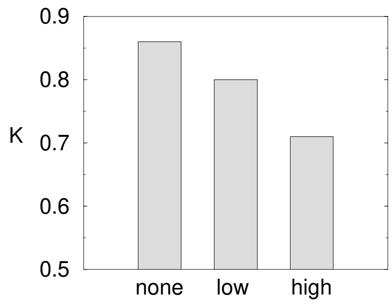


Figure 4.9: Study I: Effect of Self-Citation Ratio on Reproducibility

Figure 4.9 confirms that papers who quote previous own work only rarely or not at all seem to be annotated most consistently in our scheme. Subsequent analysis shows that part of this effect can indeed be attributed to a difficulty in distinguishing the categories OWN and OTHER. In the groups with no self citations or a low self citation ratio, we found that reproducibility does not increase too much (from K=.86 to K=.90 and from K=.8 to K=.83) if OWN and OTHER are collapsed, indicating that this distinction is not too difficult. In the high self citation group, the reproducibility increase was much higher (from K=.71 to K=.85), indicating that the distinction is more difficult in this group. This might be due to the fact that papers in the first group (and to a certain degree, in the second group) are structured in a simpler way, i.e., they might report on some isolated piece of research. However, there might be other reasons why the own new work is well-distinguished from other and own previous work in these cases.

2. There is also a difference in reproducibility between papers from different *conference types*. Out of our 21 papers, 4 were presented in student sessions, 4 came from workshops and the remaining 13 were main conference papers. Figure 4.10 shows that student session papers are the easiest to annotate, which might be due to the fact that they are shorter and have a simpler structure, with fewer mentions of previous research. Main conference papers dedicate more space to describing and criticizing other people's work than student or workshop papers (on average about one fourth of the paper). They seem to be more carefully prepared than workshop papers (and thus easy to annotate); conference authors must express themselves more clearly because they are reporting

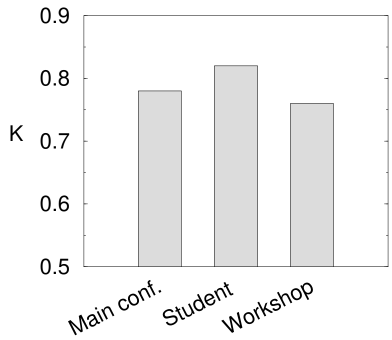


Figure 4.10: Study I: Effect of Conference Type on Reproducibility

finished work to a wider audience.

3. Another persistent problem in some papers was the distinction between OWN and BACKGROUND. This could be a sign that the authors of these papers aimed their writing at an expert audience, and thus thought it unnecessary to signal clearly which statements are commonly agreed in the field, as opposed to their own new claims. If a paper is written in such a way, its understanding requires a considerable amount of domain knowledge, which our annotators did not necessarily have. The problem here seems to be the same that Manning (1990) reports for human abstractors: the production of informative abstracts is difficult, because one needs to contrast the findings of the text with the already-established findings in the field. The recognition of the scientific contribution of a given paper requires a lot of domain knowledge in the field, particularly if it is not signalled well in the paper.

4.3.3. Study II

The only difference introduced in study II is the use of the full annotation scheme instead of the basic one.

4.3.3.1. Method

Subjects: The same annotators as in study I participated in this study.

Materials: In principle, the materials for study II were similar to the materials in study I (cf. figure 4.11). They consisted of 30 chronologically adjacent papers (papers 38–67). Papers were excluded if the first author was already represented in the materials for the given study (this was the case for papers 54, 55, 57). 5 papers were chosen as training material (papers 38, 39, 50, 51, 62). During the annotation phase, another paper turned out to be a review paper; as before, we discarded this paper from the analysis. And finally, in order to compare the performance of the tasked-untrained annotators to be used in study III to our task-trained annotators, we needed their judgement on the materials chosen for study III (papers 4 and 14). This resulted in 23 papers for annotation. For the stability experiment, we randomly chose 7 papers out of these 23.

Type of Material	Paper numbers	Sent.
Training material	38, 39, 50, 51, 62	784
Annotation material	4, 14, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 52, 56, 58, 59, 60, 61, 63, 64, 65, 66, 67	3449
Intra-annotator material	14, 41, 43, 44, 52, 58, 65	1091

Figure 4.11: Study II: Materials

Procedure: Training and annotation procedure was as in study I, except that the annotators were asked to annotate with the full annotation scheme, repeated in figure 4.12. Again, annotators were asked to annotate abstracts as well as all sentences in the document, but not acknowledgement sentences.

BACKGROUND	Generally accepted background knowledge
OTHER	Specific other work
OWN	Own work: method, results, future work...
AIM	Specific research goal
TEXTUAL	Textual section structure
CONTRAST	Contrast, comparison, weaknesses of other solution
BASIS	Other work provides basis for own work

Figure 4.12: Study II: Overview of Full Annotation Scheme

The written instructions for that scheme are reproduced in appendix C.2; they are 20 pages long. As the main decision criterion, they contain the decision tree discussed in section 3.3 (figure 3.21; p. 110). No special instructions about the use of cue phrases were given, although some of the example sentences given in the guidelines contained cue phrases.

The annotators already knew three of the seven categories from study I, and this might might have sped up the learning process with respect to completely untrained annotators; however, as there was a gap of several weeks between the two experiments, it is unlikely that this advantage was substantial.

4.3.3.2. Results and Discussion

The annotation scheme is stable (K=.82, .81, .76 for all three annotators; N=1091, k=2) and reproducible (K=.71, N=3449, k=3). Because of the increased cognitive difficulty of the task in comparison to study I, the decrease in stability and reproducibility is acceptable. Annotation between annotators varies only minimally: K=.70 (A, B); K=.70 (A, C) and K=.72 (B, C).

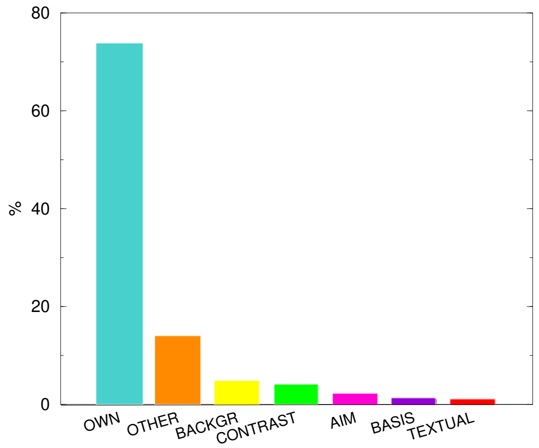


Figure 4.13: Study II: Overall Frequency of Categories

Figure 4.13 shows the relative frequencies of all seven categories. The transition between the basic categories OWN, OTHER and BACKGROUND on the one hand, and the "non-basic" categories AIM, TEXTUAL, CONTRAST and BASIS on the other is not as pronounced as we expected.

Again, variability in reproducibility is large (cf. figure 4.14), as it was in study I. Even more so than in study I, there seems to be a bimodal distribution: there is a cluster of papers with high reproducibility (K in the range of .85), and another cluster of papers with medium reproducibility (K in the range of .6). Similar explanations for this divergence as in study I are true here too: confusion between current and own previous work can be measured by self-citation ratio (cf. figure 4.15), and conference type is a predictor of overall reproducibility (cf. figure 4.16).

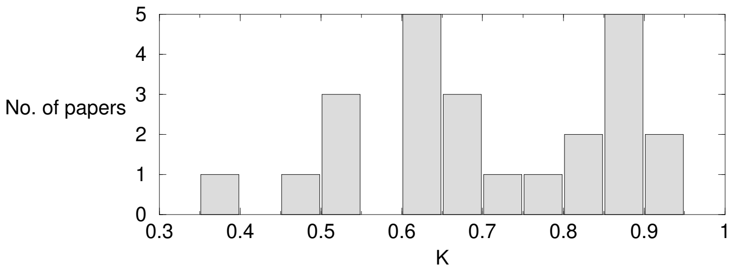


Figure 4.14: Study II: Distribution of Reproducibility Values

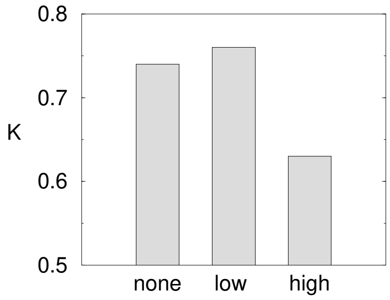


Figure 4.15: Study II: Effect of Self-citation Ratio on Reproducibility

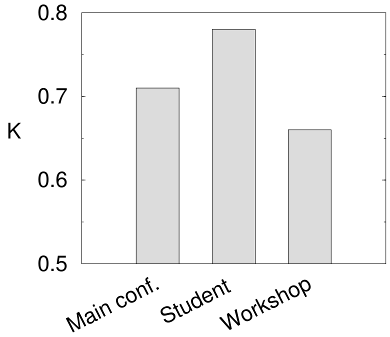


Figure 4.16: Study II: Effect of Conference Type on Reproducibility

There are problems which are specific to the new categories: annotators sometimes find it hard to distinguish neutral descriptions of other work (OTHER) from descriptions of other work which express author stance (CONTRAST and BASIS). Often, contrastive stance was not expressed openly (cf. MacRoberts and MacRoberts's (1984) explanation for this phenomenon in section 3.2.2); in order to decide if a sentence was of category BASIS, annotators needed to interpret possible reasons for the positive evaluation of other work.

AIM sentences caused the annotators problems in some cases; it can be difficult distinguishing sentences describing general aims in the field from the specific goals of a paper. All annotators perceived TEXTUAL sentences as the category which was easiest to annotate.

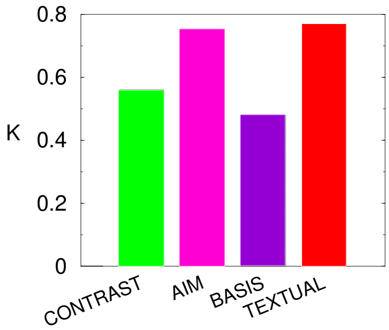


Figure 4.17: Study II: Diagnostics, Non-Basic Categories

Figure 4.17 reports how well the four non-basic categories were distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions. When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEXTUAL. This is an important result: AIM sentences constitute the single most important category in our scheme as they provide the best characterization of the research paper in a document retrieval context. Annotation performance on AIM sentences can be compared to results of free-selecting experiments where subjects were asked to identify "most relevant" sentences from a paper; traditionally, low agreement is reported for such tasks (Rath et al., 1961).

The annotators were less good at determining BASIS and CONTRAST. In section 3.2.5, we saw that there is large variation in the syntactic realization of meta-discourse signalling categories such as BASIS and CONTRAST, which makes it harder to find them. Anther reason might have to do with the location of those types of sentences in the paper: whereas AIM and TEXTUAL are usually found at the beginning or end of the introduction section, CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, BASIS and CONTRAST are more exposed to lapses of attention during annotation.

If high reliability was our priority, the annotation scheme could be simplified by creating a new category which collapses CONTRAST, OTHER and BACKGROUND. This would cause the reproducibility of the scheme to increase to K=.75. Structuring our training set in this way seems to be an acceptable compromise for our task as such a scheme would maintain most of the distinctions contained in the basic annotation scheme, while also categorizing AIM, TEXTUAL and BASIS sentences.

Figure 4.18 shows the confusion matrix between two annotators. The diagonal shows the decisions in which they agree, all other cells show decisions where they disagree. The confusion matrix is another tool apart from Krippendorff's diagnostics for detecting weaknesses in annotation schemes. One can see that the only category that AIM sentences are confused with are OWN sentences—what both categories have in common is that they describe own work. The decision of whether or not to assign an AIM label to such a sentence is a type of relevance judgement. CONTRAST sentences are often confused with OWN sentences. This is natural, as contrast sentences often compare own and other work: annotators have to judge which aspect (own or other) is more dominant, which can be hard in some cases. BACKGROUND sentences are confused with OTHER and OWN sentences, as discussed above; we suspect that the confusion with CONTRAST sentences occurs when a failure of some general method in the field is discussed. Confusion between OTHER and CONTRAST is often due to different judgement of author stance vs. neutrality expressed in the sentences. BASIS sentences are most likely to be confused with either OTHER sentences (author stance vs. neutrality), or with OWN sentences, when the annotators disagree as to if an aspect of the own work has been contributed by prior work or is first described in the current article. Appendices B.5 and B.6 show the example paper annotated by Annotators A and B; the previously shown figure 3.23 (p. 113) actually gives Annotator C's annotation of the example paper.

Figure 4.19 shows how well one annotator can predict another annotators' choice of non-basic categories. Taking Annotator B's decisions of a certain category as gold standard, recall reports how many of those instances Annotator C found, and precision reports how many of the instances that Annotator C categorized as that category, really turn out to be of that category (by Annotator B's judgement). That is, precision measures how confident we can be with the result set, whose size is measured by recall.

Annotator C achieves a precision and recall of almost 80% on TEXTUAL sentences, and 72% precision and 56% recall for AIM sentences. These values are much higher than similar values reported in earlier results for overall relevance (Rath et al., 1961). We believe that our task, given detailed guidelines, is indeed easier and better delineated than the direct determination of globally relevant sentences.

		Annotator B							
		AIM	CTR	TXT	OWN	BKG	BAS	OTH	Total
Annotator C	AIM	35	2	1	19	3		2	62
	CTR		86		31	16		23	156
	TXT			31	7			1	39
	OWN	10	62	5	2298	25	3	84	2487
	BKG		5		13	115		20	153
	BAS	2			18	1	18	14	53
	OTH	1	18	2	55	10	1	412	499
	Total	48	173	39	2441	170	22	556	3449

Figure 4.18: Study II: Confusion Matrix between Annotators B and C

	AIM	CTR	TXT	OWN	BKG	BAS	OTH
Precision	72%	50%	79%	94%	68%	82%	74%
Recall	56%	55%	79%	92%	75%	34%	83%

Figure 4.19: Study II: C's Precision and Recall per Category if B is Gold Standard

4.3.4. Study III

Study III uses a different subject pool than studies I and II. The annotators used here are not acquainted with our scheme; they are only given some general descriptions about the semantics of the categories.

4.3.4.1. Method

Subjects: 18 subjects with no prior annotation training were chosen for the second experiment. All of them have a graduate degree in Cognitive Science, with two exceptions: one was a graduate student in Sociology of Science, and one holds a master degree in English and Spanish Literature. It can be assumed that all the subjects are used to reading scientific articles, in the course of their daily work or studies, though the non-Cognitive Scientists might have come across less technical articles.

Materials: We randomly chose three papers (papers 4, 14 and 52) out of the pool of those papers for which our trained annotators had previously achieved good agreement in study I or in study II (at least K=.65). The reasoning behind this was that the task seemed cognitively difficult considering the lack of training, so we wanted to give our annotators less controversial materials. One of the three papers (paper 14) had previously resulted in much lower reproducibility (K=.67,N=205) than the other two (K=.85,N=192 for paper 4; K=.87,N=144 for paper 52).

Procedure: Each annotator was randomly assigned to a group of six, all of whom independently annotated the same single paper: group I annotated paper 4, group II paper 14 and group III paper 52. Subjects were given minimal instructions (1 page; appendix C.3), and the decision tree in figure 3.21 (p. 110).

4.3.4.2. Results and Discussion

The results show that reproducibility varies considerably between groups (K=.49, N=192, k=6 for group I; K=.35, N=205, k=6 for group II; K=.72, N=144, k=6 for group III). As Kappa is designed to abstract over the number of annotators, lower reliability in study III as compared to studies I and II is not an artifact of how K was calculated.

We must conclude that our very short instructions did not provide enough information for consistent annotation; some subjects in groups I and II did not under-

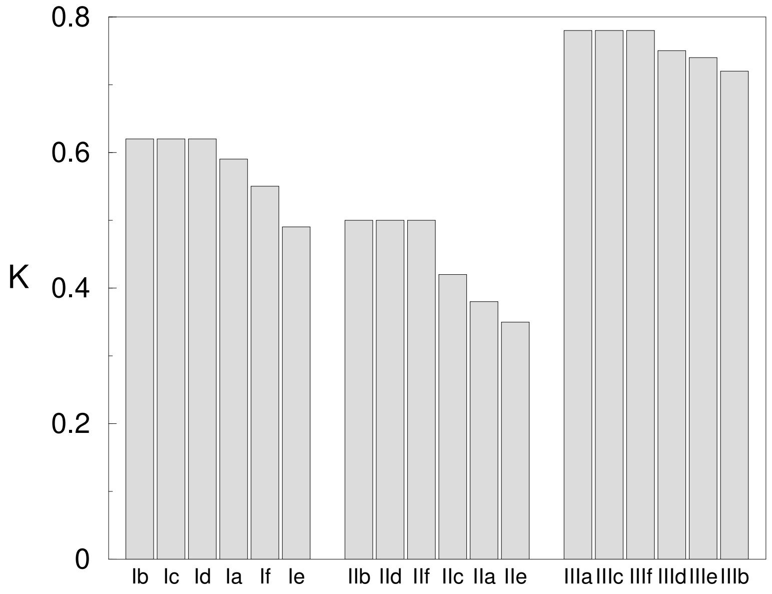


Figure 4.20: Study III: Reproducibility per Group and per Subject

stand the instructions as intended. Part of the low reproducibility results in group I and group II was due to a misunderstanding at a very superficial level. Many subjects misinterpreted the semantics of the TEXTUAL category as including sentences that refer to figures and tables in the text. This misunderstanding is easily rectifiable for future experiments, but still decreased the reliability values in this experiment considerably.

Part of the low reproducibility result can be attributed to the papers themselves: group III, which annotated the paper found to be most reproducible in study II, performed almost as well as trained annotators; group II, which performed worst, also happened to have the paper with the lowest prior reproducibility.

Figure 4.20 shows reproducibility for the most similar three annotators in each group, successively adding the next similar annotator to the pool. We can see that the performance between subjects varies much more in groups I and II than in group III, where all annotators performed more or less similarly well. Within each group, there is a subgroup of “more similar” annotators. In groups I and II, the most similar three annotators reached a respectable reproducibility (K=.63, N=192, k=3 for group I; K=.5, N=205, k=3 for group II). This result, in combination with the good performance of group III, seems to point to the fact that the annotators did have at least *some* shared understanding of the meaning of the categories.

The two subjects in study III who had no training in computational linguistics (subjects Ia and IIa) performed reasonably well: although they were not part of the circle of the most similar three subjects in their groups, their annotation also was not the odd one out.

4.3.5. Significance of Reliability Results

The reproducibility and stability values for Argumentative Zoning measured in these studies do not quite reach the levels found for, for instance, the best dialogue act coding schemes (around K=.80). Our annotation requires more subjective judgements and is possibly more cognitively complex. The reproducibility and stability results achieved with trained annotators are in the range which Krippendorff (1980) describes as giving marginally significant results if two coded variables were correlated. Of course, our requirements are rather less stringent than Krippendorff's because our annotation involves only *one* variable. On the other hand, annotation is expensive enough that simply building larger data sets is not an attractive option. Overall, we find the level of agreement which we achieved acceptable.

The single most surprising result of the experiments is the large variation in reproducibility between papers. Intuitively, the reason for this are qualitative differences in individual writing style—annotators reported that some papers are better structured and better written than others, and that some authors tend to write more clearly than others. It would be interesting to compare our reproducibility results to independent quality judgements of the papers, in order to determine if our experiments can indeed measure the clarity of scientific argumentation.

We are particularly interested in the question if shallow (human and automatic) information extraction methods, i.e. those using no domain knowledge, can be successful in a task such as Argumentative Zoning. The experiments reported in this chapter were in part conducted to establish an *upper bound* for the automatic simulation of the task. We believe that argumentative structure has enough reliable linguistic or non-linguistic correlates on the surface—physical layout being one of these correlates, along with linguistic indicators like *"to our knowledge"* and the relative order of the individual argumentative moves. The fact that the two non-computational linguists in the subject pool performed reasonably well is remarkable as the strategy that they must have used for Argumentative Zoning could not have included any domain knowledge. This result fits in nicely with the reasoning behind our approach: the implementation of Argumentative Zoning introduced in the next chapter is based on our belief that it should be possible to detect the line of argumentation of a text in a shallow, robust way.

In the framework of constructing practical gold standards for our task, the results of study II are positive as they tell us that training material gained by our method of human annotation is in principle reliable. With respect to a reduction of the effort for producing the gold standards, the outcome of study III was disappointing, as it implied that the effort cannot be reduced by simply shortening the training procedure drastically. One of the two post-analyses reported in the next section looks at a different way to reduce the effort. It determines the effect of a reduction of the textual material in each paper which is annotated. The other post-analysis looks at the argumentative structure of the author-written abstracts.

4.4. Post-Analyses

After the reliability studies had reconfirmed that the annotation can in principle be done reliably by trained annotators, Annotator C annotated the rest of the corpus. This annotation is used as system training material in chapter 5, and it also serves for the two post-analyses reported here.

4.4.1. Argumentative Structure of Author Abstracts

We wanted to establish to what extent the author abstracts differed with respect to their rhetorical structure. We therefore looked at different compositions of abstracts in terms of argumentative zones.

In the 80 papers, we found 40 different patterns, 28 of which were unique. Figure 4.21 lists all non-unique argumentative patterns in the abstracts of our corpus. The large variability reconfirms our suspicion in section 4.1.2.2 that the authors did not use a common building plan when they wrote their abstracts, in sharp contrast to how professional abstracts write their abstracts (Liddy, 1991). The composition of author abstracts seems a matter of individual choice.

The combination AIM – OWN is the single most prototypical argumentative structure we found. 29% of the abstracts in our corpus consist of this pattern. Such an abstract gives the main goal of the paper, typically followed by more detailed information about the solution. But the AIM – OWN pattern also appears as part of other abstracts: 73% of all abstracts contain it in direct sequence, and an additional 8% contain it interrupted by one other argumentative zone. A reason for the predominance of

Abstract structure	Count
AIM – OWN	23
BACKGROUND – AIM – OWN	6
OTHER – AIM– OWN	3
AIM – CONTRAST – OWN	3
OTHER – CONTRAST – AIM	3
OTHER – AIM	2
AIM – OWN – CONTRAST	2
AIM – OWN – AIM	2
AIM – OWN – BAS – OWN	2
BACKGROUND – CONTRAST – AIM – OWN	2
OWN – AIM – OWN	2
BACKGROUND – AIM	2

Figure 4.21: Typical Abstract Structures

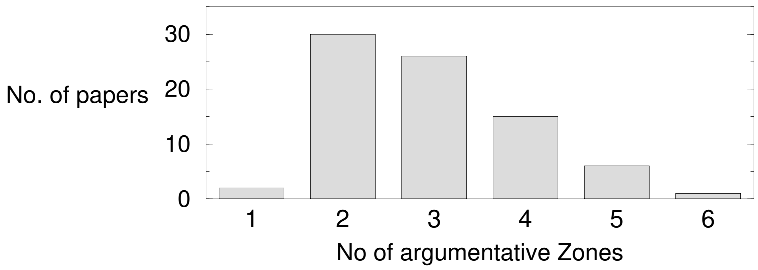


Figure 4.22: Distribution of Number of Argumentative Zones in Abstracts

this pattern might be found in the communicative function of the abstract: it is important for the success of a scientific article that the knowledge claim be established in clear terms at the earliest point of contact with the reader. This also explains the low frequency of zones referring to other researchers' work in the abstract.

AIM sentences on their own have an important function in the abstract; only one of our abstracts does not contain any AIM sentences.

Another phenomenon concerns the length of the abstracts. The average number of sentences per abstract is 4.5; the average zone in the abstract is only 1.5 sentences long. The distribution of abstract length, measured in number of argumentative zones, is given in figure 4.22. Most abstracts contain only 2 or 3 argumentative zones (average: 2.95). That is, the author abstracts in our corpus do not cover enough argumentative zones to be useful for document characterization, apart from the fact that their structure is very heterogeneous. This reconfirms our hypothesis from section 4.1.2.2: author abstracts do not provide good gold standards for Argumentative Zoning.

4.4.2. Reduction of Annotation Areas

Annotating texts with our scheme is time-consuming, so we wanted to test if the annotation of only *parts* of the source texts (which would certainly increase efficiency) would still result in reliable hand-annotated training material.

In general, we expect most of the non-basic categories (which carry the most information for our task) to be located in the periphery of the paper. For example, the TEXTUAL zone makes most sense at the end of the introduction. If an introduction section is rich in non-basic categories, it probably displays a miniature argumentative structure of the whole paper, which is generally held to be a good strategy for writing introductions (Swales, 1990; Manning, 1990). Similarly, the abstract and conclusions of source texts are often considered as "condensed" versions of the contents of the entire paper. It is thus plausible that these sections could contribute more "important" sentences to the gold standard. Additionally, one could expect these areas to be amongst the most clearly written and information rich sections in a paper.

In the following study, sections entitled *Motivation*, *Background* or *Summary* are treated as if they were called *Introduction* or *Conclusions*, respectively. As *Discussion* sections contain more speculative material, we do not treat them like *Conclusions*. Many papers do not contain explicit rhetorical sections, so we also report values for approximations of these sections: the first and last one fifth (and one tenth/twentieth) of the paper.

The abstract has a special status. As it is not clear if the abstract itself would be available for extraction in a typical practical scenario, we also report results for *aligned* abstract sentences, as discussed in section 4.1.2.2.

We test the hypothesis that the reproducibility in these special areas is higher than the overall reproducibility. If it turned out to be the case, we could either reduce annotation to these areas, or use sentences from those areas as "best fillers" to a slot (cf. section 4.1.3).

Results are given in figure 4.23: only some of the supposedly "good" areas for annotation restriction show an increase in reliability, namely only *Abstract* and *Conclusions*. These two sections have the clearest summarization function of the entire article. The effect that abstracts are more consistently annotated is even stronger in the

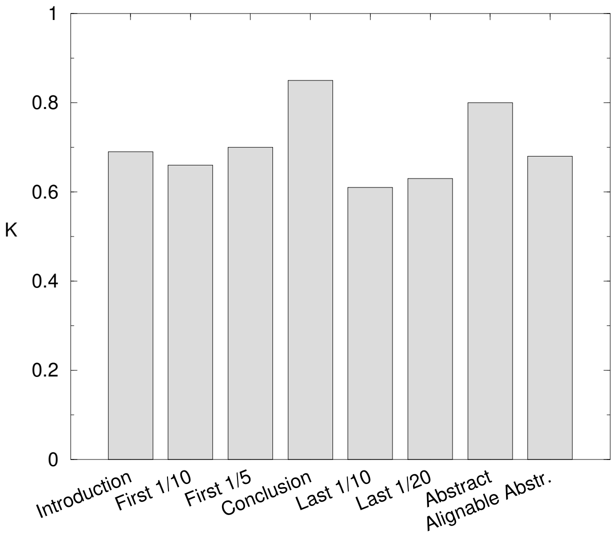


Figure 4.23: Reproducibility by Areas

basic scheme (not shown here): reproducibility within abstracts shows the very high value of K=.92. This means that authors make particularly clear in the abstract what their own contributions are.

All other areas actually show a *lower* reproducibility than the average. This is true in particular for the areas defined by absolute location (e.g. the last 1/20). These areas are therefore *not* a good approximation to *Conclusions* type material. It looks as if the last few lines in papers that do not have an explicitly marked conclusion section should not be considered at all—these sentences do not contribute "summary" type information. The *Introduction* section shows a slight decrease in reproducibility, and location approximations of introduction sections also perform badly. Reproducibility is considerably lower in alignable abstract sentences than in the abstract itself. This is consistent with our observation in section 4.1.2.2 that the rhetorical status of the aligned abstract sentences is often different from the status of the corresponding document sentences.

But there is a second point we have to take into account when restricting the areas for gold sentence selection: it is also necessary to cover all argumentative categories, as discussed in section 4.1.3. Obviously, any strategy of annotation restriction will give us fewer gold standard sentences per paper, so it is an empirical question whether there are still enough candidate sentences for all seven categories.

Some documents do not even contain all argumentative zones. In our data, each document contains at least one AIM sentence (this is required in the guidelines);

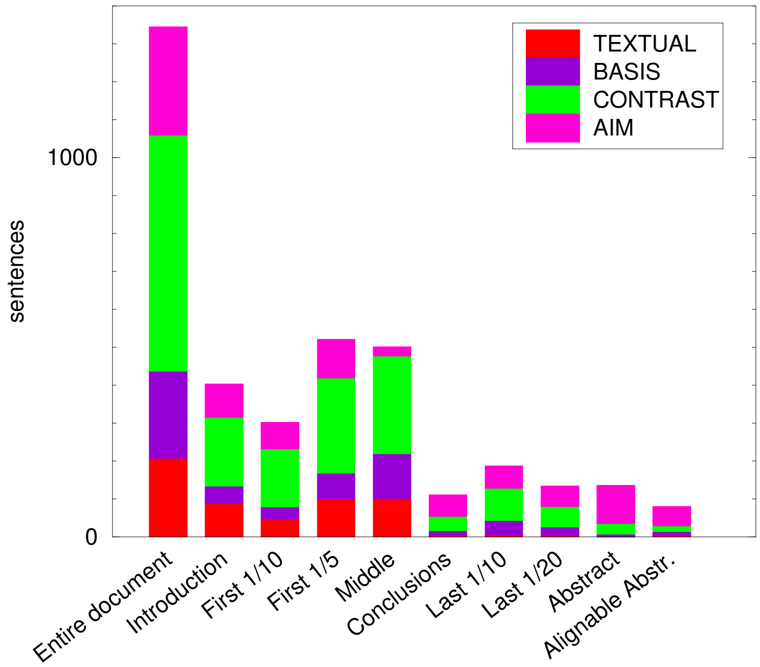


Figure 4.24: Non-Basic Areas by Categories; Absolute Values

almost every document contains at least one CONTRAST sentence (3 documents do not, i.e. 4% of our corpus). However, the use of TEXTUAL zones seems to depend much more on personal writing style. 26 % of the documents do not contain TEXTUAL zones. As the papers are conference papers and thus rather short, authors did not always perceive the function of explicitly previewing the textual presentation as necessary. Similarly, BASIS sentences are not present in 20% of the papers. However, the presence or absence of BASIS sentences seems to have less to do with writing style and more with the type of research done.

The values in figure 4.24 show absolute numbers for the occurrence of non-basic categories in special areas. For example, we can see that there are not many alignable abstract sentences anywhere in the document—a gold standard defined by alignable sentences only would thus result in bad *overall* coverage, as we have argued in section 4.1.2.2.

Figure 4.25 shows which categories can be found in a given area, and figure 4.26 shows in which areas a given category can be found. We see that some areas show a particularly low *variability* with respect to categories. Conclusions, for example, mainly consist of OWN sentences, with occasional AIM and CONTRAST sentences. Conclusions capitalize on the overall research process: they highlight own

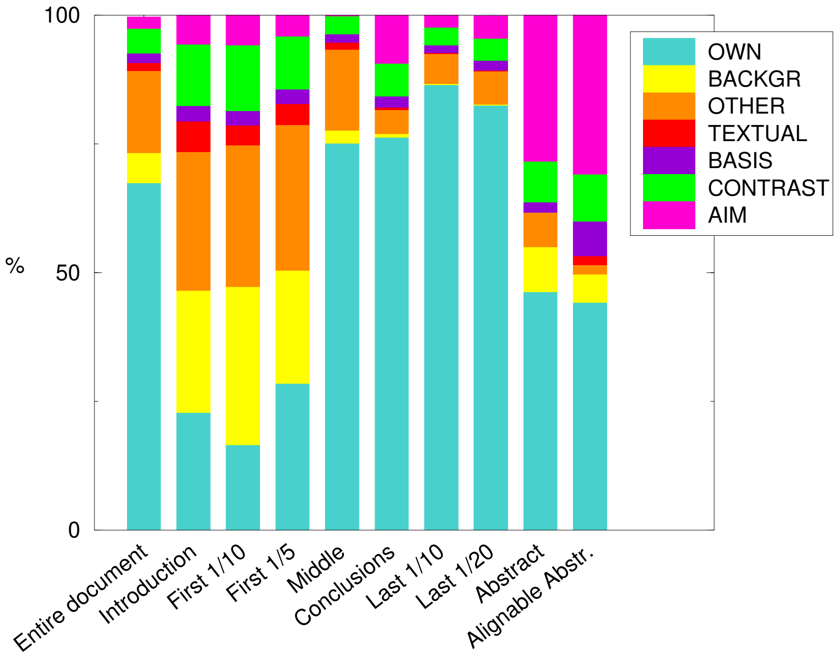


Figure 4.25: Areas by Categories; Relative Values

contribution, relevance of results, limitations, future work, and advantages over rival approaches. For some tasks, this type of information might be enough; however, we predict that it would not be enough for ours.

The relatively high proportion of AIM sentences found in abstracts would be advantageous for our task. However, even if we considered conclusion and (alignable) sentences together, coverage would still be low for certain categories, e.g. BACKGROUND, BASIS and TEXTUAL. All of these categories can be found in the introduction. It is the variety of argumentative categories in the introduction which makes annotation of this section more difficult (cf. the comparatively low reproducibility in figure 4.23), but also more rewarding for our task.

A compromise between time efficiency and quality is to annotate abstracts, introductions and conclusions where available, and first and last paragraphs as a fall back option. The price to be paid for this efficiency is in coverage and comparability. Annotated material occurring in the large area marked "Middle" or "Rest" (all document areas except alignable sentences, introduction and conclusions; black in figure 4.26), including BASIS, would get lost. Also, we cannot be sure that a given paper is written in a modular way, i.e. that it reiterates important material from the middle of the document in the periphery—some do not repeat information introduced from the abstract in the introduction section (cf. section 4.1.2.2). This is another reason why the quality

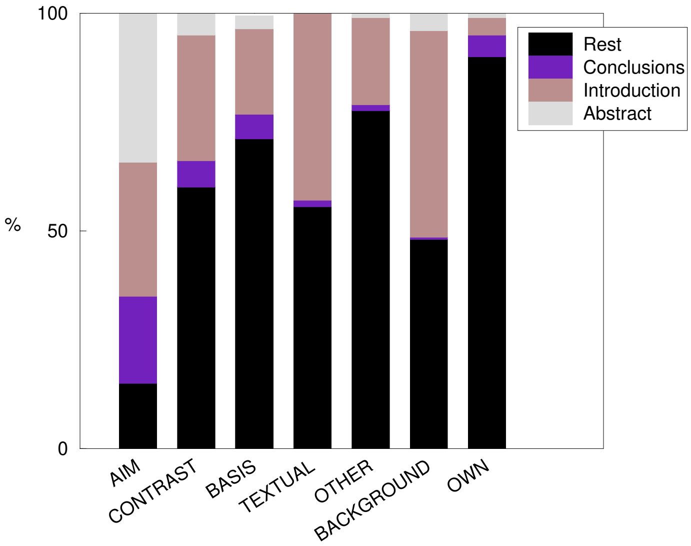


Figure 4.26: Categories by Areas

of area-reduced annotation might be lower than unrestricted annotation.

In sum, the annotation effort can be reduced by restricting the annotation to certain areas within a paper, but such a restriction has its price in quality of the gold standards. One could restrict the annotation to sentences appearing in the introduction section, even though annotators will find them harder to classify, or to all alignable abstract sentences, even if there are not many of them overall, or to conclusion sentences, even if the coverage of different argumentative categories is very restricted. The implications for Argumentative Zoning gold standards are that the advantage of time savings have to be weighed against task considerations in the concrete scenario.

4.5. Conclusion

In the first section of this chapter we discussed the question how a practical gold standard for a task like Argumentative Zoning could be constructed, and how its value could be evaluated. This discussion led to a list of desired properties of a gold standard—some of which are difficult to achieve with a surface-based evaluation strategy like ours. We have discussed why simpler gold standards, such as targets keys and free-selected sentences, are not sufficient in our text type and task. In particular, we have argued that similarity with abstract sentences does not automatically constitute a good gold standard; evidence presented in section 4.4.1 confirms this argument. Our methodology for arriving at a gold standard relies on human judgements of every sentence in the document. We decided to conduct reliability studies to measure the degree of human agreement on the task.

In section 4.2, we advocate the Kappa coefficient as a measure for annotation similarity. The main part of the chapter (section 4.3) presents the experiments: they demonstrate that the annotation scheme can indeed be learned by trained annotators and subsequently applied in a consistent way. In particular, study I shows that the basic annotation scheme, which distinguishes sentences on the basis of attribution of scientific authorship, is particularly reliable, both over time as well as between annotators. This is important, as the concept of intellectual attribution is new and central to our model of argumentation (cf. section 3.2).

Study II examines Argumentative Zoning (i.e. it uses the full annotation scheme). It shows that the two most important additional categories, AIM and TEXTUAL, are annotated reliably, but we identified some minor difficulties with the two categories BASIS and CONTRAST. As the reliability of the full scheme (as used in study II) is still acceptable, we decided to use the annotated corpus as our gold standard. This corpus is to be used for training an automatic Argumentative Zoning system, and also for intrinsic evaluation.

Study III tentatively confirms the intuitivity of the categories of the scheme, but also shows that Argumentative Zoning is a complex task which requires a certain training period in order to be performed consistently. In particular, our results show that very short annotation instructions do not provide enough information for Argumentative Zoning.

In section 4.4.1 we report the results of two post-analyses. One looks at the argumentative zones found in author abstracts and reconfirms that they cannot be directly used as gold standard. The other investigates the possibility of restrictions of the practical annotation effort by annotating only parts of papers. Our hypothesis that the reliability of the annotation in special areas of the paper would be higher in comparison to the reliability achieved overall has not been confirmed in all cases. The best gold standard is achieved when the entire paper is annotated, though we have given some alternatives for cases when such annotation might seem too costly.

Chapter 5

Automatic Argumentative Zoning

In this chapter, we will describe one method for solving the task of Argumentative Zoning automatically. As previously detailed, the task is to determine the best argumentative category for each sentence, out of a fixed list of seven categories. We have already discussed how we collected human judgements about the argumentative category for each sentence in our corpus. In this chapter, we will report on a prototype system which, on the basis of algorithmically determinable features of the sentence, learns the correlation between the human judgements and the features. An alternative system determines argumentative zones in a rule-based way. In the following, we will give an overview of the definition of the features and of the implementation, followed by results of an intrinsic evaluation.

5.1. Overview of Automatic Argumentative Zoning

Figure 5.1 gives an overview of the processes involved in automatic Argumentative Zoning. Before the experiment, the following steps had to be performed:

- *1. Feature definition*: Sentential features had to be determined which we expect to correlate with argumentative status. It is important that these features can be easily determined automatically. Our choice of features is described in section 5.2.

- *2. Human annotation*: As already discussed, a gold standard is needed, in our case in the form of human annotation of argumentative categories (cf. 4). The annotation is used for training and for evaluation.

The statistical system consists of a training and a testing phase. During training, the following steps are performed:

- *3. Preprocessing*: Each document in the training corpus is preprocessed into a machine readable format with minimal mark-up, e.g. divisions and headlines are marked (cf. section 5.3.2).
- *4. Feature determination*: For each sentence in the training corpus, values for each of the sentential features are determined automatically (cf. section 5.3.3).
- *5. Statistical training*. Several statistical classifiers are used for statistical model building, determining the correlation between sentential features and argumentative zones (cf. section 5.3.4).

Testing, i.e. the application of the statistical model to a new (test) document, uses preprocessing and feature determination in the same way as during training. This is followed by a step of

- *6. Statistical classification*: Using the model acquired in the training phase, each sentence is classified by its most likely argumentative status.

Alternatively, there is also a different system for Argumentative Zoning:

- *7. Symbolic rules*: These rules operate on the representation derived in the feature determination step (cf. section 5.3.5).

We compare human-annotated test documents against the output of the symbolic and the statistical Argumentative Zoning systems in the evaluation:

- *8. Intrinsic Evaluation:* Some parts of the training corpus are singled out for testing (i.e. they are *not* used for training). The system output is then compared with the human classification (cf. sections 5.4.1 and 5.4.2).

Finally, the output of the systems has to be displayed:

- *9. Postprocessing*: The output of the automatic and the human annotation, and the output of the automatic feature determination, are transformed into HTML (using cascading style sheets) so that the paper plus all of its annotation can be displayed in an HTML browser, eg. Netscape.

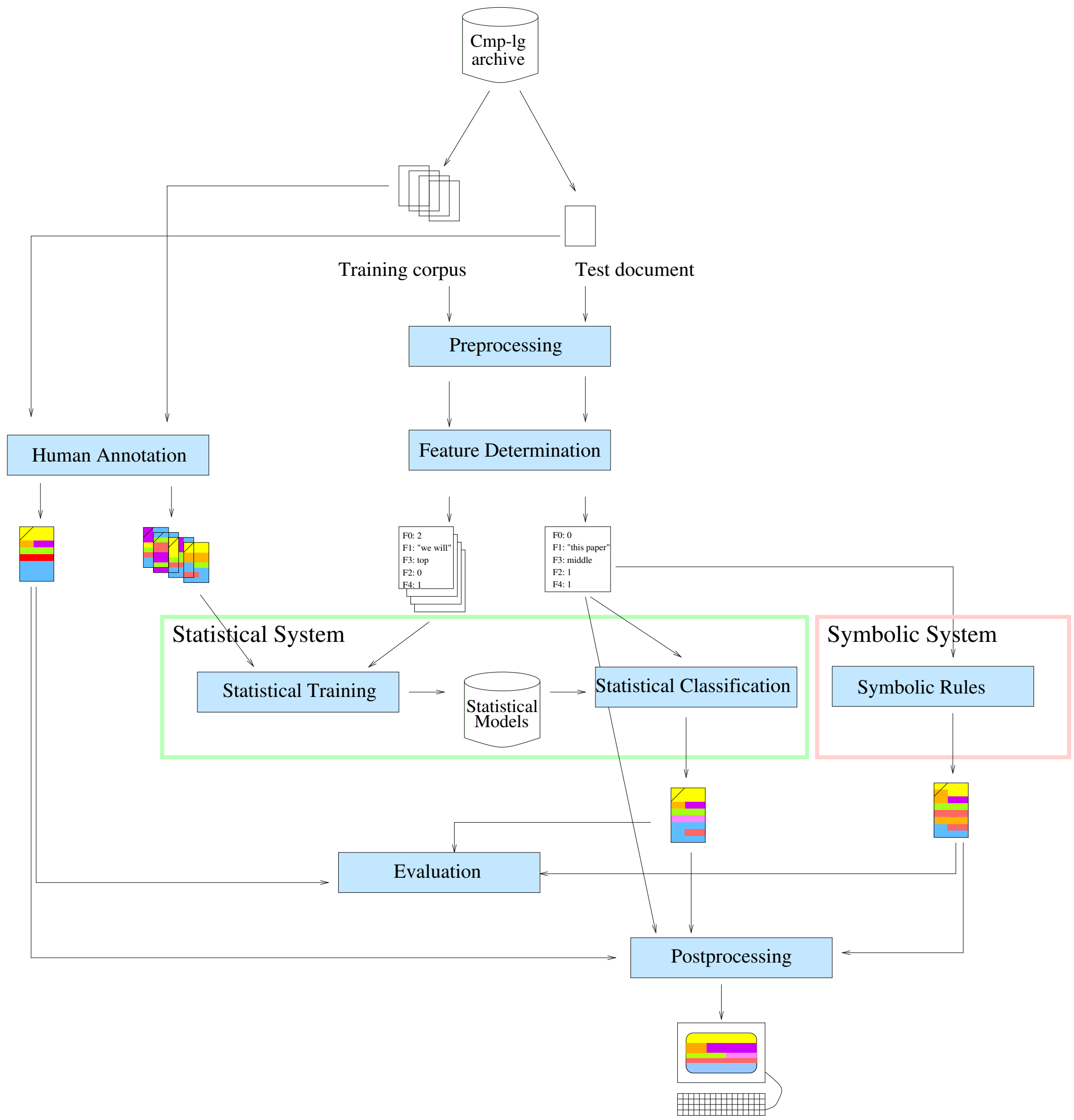


Figure 5.1: Overview of our Implementation of an Argumentative Zoner

Another overview of this rather complex setup is given in figure 5.2, which concentrates on the representations of the corpus at different stages of processing. The documents are taken from the source archive in two formats (LaTeX and PostScript). The PostScript versions are printed out and hand-annotated, the corresponding LaTeX versions are converted into XML. They constitute the training material for automatic Argumentative Zoning. After the training corpus has been automatically annotated, intrinsic evaluation is measured by the Kappa statistics, and postprocessing produces web-browsable HTML representations of the output of seen and unseen papers.

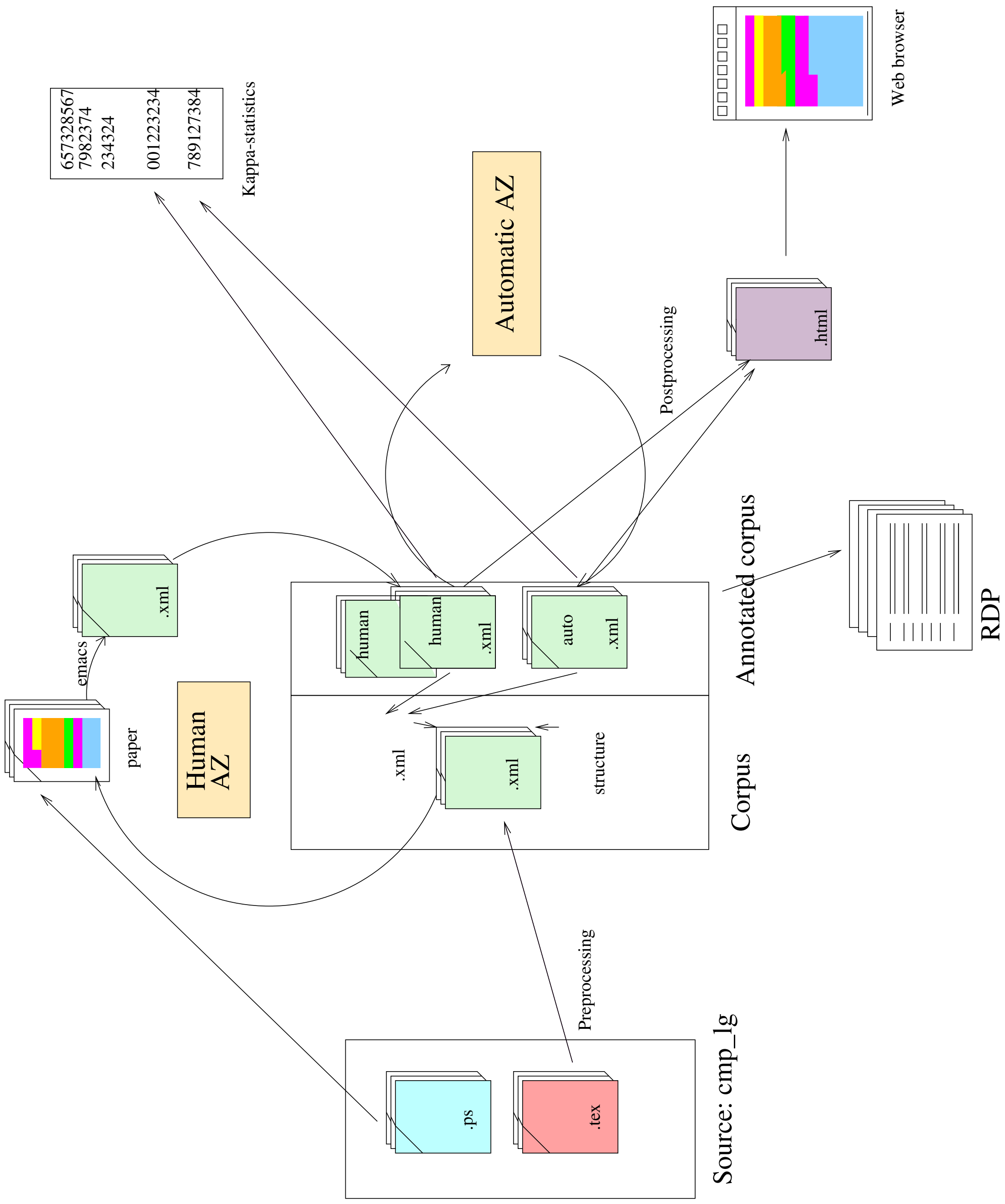


Figure 5.2: Data Flow in the Argumentative Zoner

5.2. Correlates of Argumentative Status

The argumentative status of a sentence is a property that is too difficult to determine directly algorithmically. Instead, we define heuristics which measure how appropriate it is to assign a given argumentative zone to a sentence. For this end, we need to define operationally tractable correlates (sentential features) which capture some characteristic aspect of that sentence's argumentative status.

It is generally assumed that appropriate correlates exist for similar tasks. For example, human summarizers are guided by sentential features like location and the occurrence of certain cue phrases when they determine importance of a textual segment (Cremmins, 1996); and the text extraction literature provides us with a pool of such features (heuristic measures) for sentence *relevance* (Paice, 1990; Luhn, 1958; Baxendale, 1958; Edmundson, 1969; Kupiec et al., 1995).

The task of Argumentative Zoning moves away from the concept of sentence relevance towards a new concept of argumentative status. Our annotation scheme can be interpreted as encoding different *types* of relevance. We have defined four different kinds of sentences which are particularly important for the global argumentation of the paper (the non-basic categories), and three categories which provide background information. All of these are important for different reasons. We have assumed so far that there are correlates of this argumentative status in our texts which can be read off the surface.

It might well be that the features which are useful for our task differ from the ones used for determining global relevance. Figure 5.3 gives an overview of our feature pool. Some of the features we use (the Content, Explicit Structure, Absolute Location, Formulaic and Sentence Length features) are borrowed from the text extraction literature, but in some cases, changes were necessary; the Formulaic feature, for example, is an elaboration of similar, simpler features used previously. We also use features not typically used for text extraction, namely the Syntactic, Citation and Agentivity Features; as far as we know, we are the first to define these for any task.

When defining the features, we tried to make them maximally *distinctive*. In order to do so, we used information provided in *contingency tables*. A contingency table lists the values of a given feature with its counts in the corpus, cf. figure 5.4.

Distinctive features have heterogeneous (skewed) distributions, i.e. distributions which differ as much as possible from the overall distribution of categories. There are statistical measures for this heterogeneity, e.g. g-score (Dunning, 1993). In section 5.3.3, we will provide the contingency tables for each of our features; the use of contingency tables for statistical classification will be discussed in section 5.3.4.

Type	Name	Feature description	Feature values
Content Features	`Cont-1`	Does the sentence contain “significant terms” as determined by the *tf/idf* measure?	Yes or No
	`Cont-2`	Does the sentence contain words also occurring in the title or headlines?	Yes or No
Absolute location	`Loc`	Position of sentence in relation to 10 segments	A-J
Explicit structure	`Struct-1`	Relative and absolute position of sentence within section (e.g. first sentence in section or somewhere in second third)	7 values
	`Struct-2`	Relative position of sentence within a paragraph	Initial, Medial, Final
	`Struct-3`	Type of headline of current section	16 prototypical headlines or *Non-Prototypical*
Sentence length	`Length`	Is the sentence longer than a certain threshold in words?	Yes or No
Verb Syntax	`Syn-1`	Voice (of first finite verb in sentence)	Active or Passive or NoVerb
	`Syn-2`	Tense (of first finite verb in sentence)	9 simple and complex tenses or NoVerb
	`Syn-3`	Is the first finite verb modified by modal auxiliary?	Modal or no Modal or NoVerb
Citations	`Cit-1`	Does the sentence contain a citation or the name of an author contained in the reference list?	Citation, Author Name or None
	`Cit-2`	Does the sentence contain a *self* citation?	Yes or No or NoCitation
	`Cit-3`	Location of citation in sentence	Beginning, Middle, End or NoCitation
Formulaic expressions	`Formu`	Type of formulaic expression occurring in sentence	20 Types of Formulaic Expressions + 13 Types of Agents or None
Agentivity	`Ag-1`	Type of Agent	13 different types of Agents or None
	`Ag-2`	Type of Action, with or without Negation	20 different Action Types X Negated/Non-negated, or None

Figure 5.3: Overview of Feature Pool

Paragraph (`Struct-2`)	AIM	BAS	BKG	CTR	OTH	OWN	TXT	Total
Initial	117	92	267	135	601	2532	73	3817
Medial	56	87	306	289	971	3779	68	5556
Final	34	47	147	172	442	2125	82	3049
Total	207	226	720	596	2014	8436	223	12422

Figure 5.4: A Contingency Table: Paragraph Feature

Another desired property is *coverage* (as opposed to *peakiness*). Some features are strong indicators of a certain category, but occur very rarely in the corpus. For the average sentence, such a feature would not be of help for classification. Moreover, such features lead to *over-fitting*, a problem which occurs when features encode idiosyncrasies of the training data which are accidental to the data. The feature will then not provide useful information for unseen, but similar data. An example for a peaky feature is the occurrence of the phrase "*in this paper*" in a sentence. Evenly distributed features (e.g. verb tense) have a higher coverage, i.e., they can be more reliably estimated from text. They typically do not give strong indications, but many of them in combination might influence the statistical classification into the right direction. We have tried to find a compromise between features that are peaky and those that are evenly distributed.

The choice of the values for the features is not independent of the classification method chosen. We initially followed Kupiec et al. (1995) in using a Naive Bayesian classifier. Later, we used other classifiers, but the original design of the features was influenced by the intention to use them in a Naive Bayesian classifier. This classifier demands that features must have discontinuous values, and in practice it also implies that feature values all fall into a small set of distinct values. Too many values might influence classification results negatively as there might not be enough training material available for the rare values. Thus, we often had to *cluster* values into classes; we did so manually. Another limitation is that Naive Bayes allows only one value of a feature per classified item. Additionally, Naive Bayes assumes that the features are *statistically independent* of each other, so we tried to identify features which would classify sentences into certain categories for reasons different from the other features in the feature pool.

5.2.1. Traditional Features

5.2.1.1. Content Features

The assumption behind the content features is that concepts (approximated by textual strings) are representations of the semantics of the text span in the context of the overall document. Different content features might differ in exactly *how* they determine the most salient concepts in a text span. Content features are used in most of today's sentence extractors, i.e. for determining global sentence relevance.

The two content features we use are different from the other heuristics in our pool in that they concentrate on *subject matter* rather than more structural or rhetoric cues. We hypothesized that content features should be less important for Argumentative Zoning than the other features, as it is not immediately obvious how the fact that a certain sentence contains characteristic subject-matter key words would help determine its argumentative category.

Term Frequency (`Cont-1`): `Cont-1` uses the *tf/idf* (term frequency times inverse-document-frequency) method, which employs lexical frequency to identify concepts that are characteristic for the contents of the document. The *tf/idf* method is successfully used for information retrieval (Salton and McGill, 1983).

tf/idf tries to identify diagnostic units (textual spans) which are frequent in one document but rare in the overall collection. This is achieved by combining the relative frequency weights (*tf*) with a function of the inverse frequency of the diagnostic unit in the overall text collection (the *idf* element), e.g. the number of documents where this term occurs, or the frequency of overall occurrences:

$$tf/idf_w = tf_w * log(\frac{100*N}{df_w})$$

td/idf_w:	td/idf weight for diagnostic unit w
tf_w:	term frequency of w in document
df_w:	number of documents containing diagnostic unit w or number of occurrences of w in document collection
N:	number of documents in collection

If a diagnostic unit appears often in the overall collection, it is assumed that it represents a concept which is common in the domain, and which has a low discriminating power—as a result, it is penalized by a low *idf* score. If a diagnostic unit appears only once, it might be noise (e.g. misspelled words); such words can be filtered out by frequency thresholds.

In the first text extraction experiments (Luhn, 1958; Baxendale, 1958), a predecessor of today's *tf/idf* formula was used, which relied only on the *tf* part. There are variations of the formula used in the literature (e.g. Brandow et al. (1995) use the logarithm also for the *tf* part). Other approaches have varied the diagnostic units used. Luhn's (1958) diagnostic units were the most frequent content word *stems* (after function words had been stripped out with a stop list), i.e. *"hypothesis"* and *"hypothesize"* were reduced to the same stem. Nowadays, the simplest implementations use either full words or lemmas (words normalized to their lexicon entries). Other implementations use nominal pairs, or noun groups determined by partial parses, derived by techniques like chunking (shallow parsing of NP and VP complexes; Abney 1990; Grefenstette 1994). Georgantopoulos (1996) improves results achieved by Finch and Mikheev (1995) by using noun groups as diagnostic units.

There has also been criticism of the method, as it cannot handle synonymy, pronominalization, general co-referentiality and conceptual generalizations such as the replacement of a list by its superordinate term (Hovy and Lin, 1999; Mauldin, 1991). This limitation has been referred to in IR as the "keyword boundary".

An additional criticism questions if the application of *tf/idf* measures from document retrieval to text extraction is sensible, i.e. if the transition from *documents* as units of scoring to smaller units like *sentences* actually works. (Hearst, 1997) voices the intuition that *tf/idf* works much better to determine important concepts which distinguish *between* documents rather than between smaller segments *within* a document:

> [...] the estimates of importance that *tf/idf* makes seem not to be accurate enough within the scope of comparing adjacent pieces of text to justify using this measure [...] (Hearst, 1997, p. 44)

Title Words (`Cont-2`): `Cont-2` draws its definition of what a good keyword is from occurrences of a word in the title and headline. This feature goes back to Edmundson (1969). The assumption is that words occurring in the title are good candidates for document specific concepts. Particularly in experimental disciplines, titles can be a document surrogate in themselves, as they often summarize the main knowledge claim of the document ("*Low Dose Dobutamine Echocardiography Is More Predictive of Reversible Dysfunction After Acute Myocardial Infarction Than Resting Single Photon Emission Computed Tomographic Thallium-201 Scintigraphy"*; American Heart Journal, 134(5): 822-834, 1997).

Along the same lines, headlines are considered summaries of the major sections of the document—unless they are prototypical headlines such as *Introduction* or *Results*.

However, in other fields, "jokey" titles have become fashionable (*"Four out of five ain't bad"*; Archives of General Psychiatry, 55(10): 865-866, 1998). This practice makes reliance on title heuristics risky as titles do not necessarily express the document's topic anymore.

5.2.1.2. Absolute Location

The next two features use the location of a sentence in text. In many previous experiments, local organization within a section has been correlated with importance. Experiments in text extraction have assumed that more relevant sentences can be found in the periphery of the document (Edmundson, 1969). Indeed, in other genres like newspaper text, location has been shown to be the single most important feature for text extraction (Brandow et al., 1995; Hovy and Lin, 1999).

Absolute location, in terms of absolute spatial organization of information in the linear medium of text, should be a good correlate for Argumentative Zoning. Readers have certain expectations of how the chain of argumentation will proceed and which argumentative components are handled in which areas of the paper.

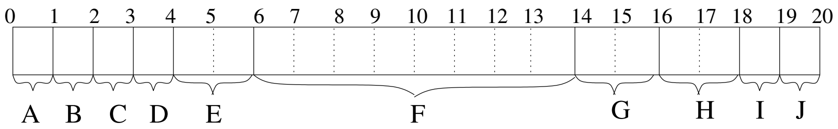


Figure 5.5: Values for Location Feature

We divide the document into 20 equally sized segments; we then collapse some of these (cf. figure 5.5), resulting in 10 differently-sized segments which mimic the structure of ideal documents. Segment size is smaller towards the beginning and the end of the document, where documents are often written more densely, i.e. where we expect the author's rhetorical units to be smaller. In the middle, the segments are larger (cf. segment F in figure 5.5, which covers 40% of the text).

5.2.1.3. Structural Correlates

The structural features seek to exploit the explicit hints given by the author about the structure of the paper.