An annotation scheme for discourse-level argumentation in research articles

Simone Teufel[‡] and **Jean Carletta**[†] and **Marc Moens**[‡]
[‡]HCRC Language Technology Group and
[†]Human Communication Research Centre
Division of Informatics
University of Edinburgh
S.Teufel@ed.ac.uk, J.Carletta@ed.ac.uk, M.Moens@ed.ac.uk

Abstract

In order to build robust automatic abstracting systems, there is a need for better training resources than are currently available. In this paper, we introduce an annotation scheme for scientific articles which can be used to build such a resource in a consistent way. The seven categories of the scheme are based on rhetorical moves of argumentation. Our experimental results show that the scheme is stable, reproducible and intuitive to use.

1 Introduction

Current approaches to automatic summarization cannot create coherent, flexible automatic summaries. Sentence selection techniques (e.g. Brandow et al., 1995; Kupiec et al. 1995) produce extracts which can be incoherent and which, because of the generality of the methodology, can give under-informative results; fact extraction techniques (e.g. Rau et al., 1989, Young and Hayes, 1985) are tailored to particular domains, but have not really scaled up from restricted texts and restricted domains to larger domains and unrestricted text. Spärck Jones (1998) argues that taking into account the structure of a text will help when summarizing the text.

The problem with sentence selection is that it relies on extracting sentences out of context, but the meaning of extracted material tends to depend on where in the text the extracted sentence was found. However, sentence selection still has the distinct advantage of robustness.

We think sentence selection could be improved substantially if the global rhetorical context of the extracted material was taken into account more. Marcu (1997) makes a similar point based on rhetorical relations as defined by Rhetorical Structure Theory (RST, (Mann and Thompson, 1987)). In contrast to this approach, we stress the importance of rhetorical moves which are *global* to the argumentation of the paper, as opposed to local RST–type moves. For example, sentences which describe weaknesses of previous approaches can provide a good characterization of the scientific articles in which they occur, since they are likely to also be a description of the problem that paper is intending to solve. Take a sentence like "*Unfortunately, this work does not solve problem X*": if X is a shortcoming in someone else's work, this usually means that the current paper *will* try to solve X. Sentence extraction methods can locate sentences like these, e.g. using a cue phrase method (Paice, 1990).

But a very similar-looking sentence can play a completely different argumentative role in a scientific text: when it occurs in the section "Future Work", it might refer to a minor weakness in the work presented in the source paper (i.e. of the author's *own* solution). In that case, the sentence is *not* a good characterization of the paper.

Our approach to automatic text summarization is to find important sentences in a source text by determining their most likely argumentative role. In order to create an automatic process to do so, either by symbolic or machine learning techniques, we need training material: a collection of texts (in this case, scientific articles) where each sentence is annotated with information about the argumentative role that sentence plays in the paper. Currently, no such resource is available. We developed an annotation scheme as a starting point for building up such a resource, which we will describe in section 2. In section 3, we use content analysis techniques to test the annotation scheme's reliability.

2 The annotation scheme

We wanted the scheme to cover one text type, namely research articles, but from different presentational traditions and subject matters, so that

we can use it for text summarization in a range of fields. This means we cannot rely on similarities in external presentation, e.g. section structure and typical linguistic formulaic expressions.

Previous discourse-level annotation schemes (e.g. Liddy, 1991; Kircz, 1991) show that information retrieval can profit from added rhetorical information in scientific texts. However, the definitions of the categories in these schemes relies on domain dependent knowledge like typical research methodology, and are thus too specific for our purposes.

General frameworks of text structure and argumentation, like Cohen's (1984) theoretical framework for general argumentation and Rhetorical Structure Theory (Mann and Thompson, 1987), are theoretically applicable to many different kinds of text types. However, we believe that restricting ourselves to the text type of research articles will give us an advantage over such general schemes, because it will allow us to rely on communicative goals typically occurring within that text type.

Swales' (1990) CARS (Creating a Research Space) model provides a description at the right level for our purposes. Swales claims that the regularities in the argumentative structure of research article introductions follow from the authors' primary communicative goal: namely to convince their audience that they have provided a contribution to science. From this goal follow highly predictable subgoals which he calls *argumentative moves* ("recurring and regularized communicative events"). An example for such a move is "*Indication of a gap*", where the author argues that there is a weakness in an earlier approach which needs to be solved.

Swales' model has been used extensively by discourse analysts and researchers in the field of English for Specific Purposes, for tasks as varied as teaching English as a foreign language, human translation and citation analysis (Myers, 1992; Thompson and Ye, 1991; Duszak, 1994), but always for manual analysis by a single person. Our annotation scheme is based on Swales' model but we needed to modify it. Firstly, the CARS model only applies to introductions of research articles, so we needed new moves to cover the other paper sections; secondly, we needed more precise guidelines to make the scheme applicable to reliable annotation for several non-discourse analysts (and for potential automatic annotation).

For the development of our scheme, we used computational linguistics articles. The papers in our collection cover a challenging range of subject matters due to the interdisciplinarity of the field, such as logic programming, statistical language modelling, theoretical semantics and computational psycholinguistics. Because the research methodology and tradition of presentation is so different in these fields, we would expect the scheme to be equally applicable in a range of disciplines other than those named.

Our annotation scheme consists of the seven categories shown in Figure 1. There are two versions of the annotation scheme. The *basic* scheme provides a distinction between three textual segments which we think is a necessary precondition for argumentatively-justified summarization. This distinction is concerned with the attribution of *authorship* to scientific ideas and solutions described in the text. Authors need to make clear, and readers need to understand:

- which sections describe generally accepted statements (BACKGROUND);
- which ideas are attributed to some other, specific piece of research outside the given paper, including own previous work (OTHER);
- and which statements are the authors' own *new* contributions (OWN).

The *full* annotation scheme consists of the basic scheme plus four other categories, which are based on Swales' moves. The most important of these is AIM (Swales' move "*Explicit statements of research goal*"), as these moves are good characterizations of the entire paper. We are interested in how far humans can be trained to consistently annotate these sentences; similar experiments where subjects selected one or several 'most relevant' sentences from a paper have traditionally reported low agreement (Rath et al., 1961). There is also the category TEXTUAL (Swales' move "*Indicate structure*"), which provides helpful information about section structure, and two moves having to do with attitude towards previous research, namely BASIS and CONTRAST.

The relative simplicity of the scheme was a compromise between two demands: we wanted the scheme to contain enough information for automatic summarization, but still be practicable for hand coding.

Annotation proceeds sentence by sentence according to the decision tree given in Figure 2. No instructions about the use of cue phrases were given, although some of the example sentences given in the guidelines contained cue phrases. The categorisation task resembles the judgements performed e.g. in dialogue act coding (Carletta et al.,

BASIC SCHEME	BACKGROUND	Sentences describing some (generally accepted) background knowledge	**FULL SCHEME**
	OTHER	Sentences describing aspects of some specific other research in a neutral way (excluding contrastive or BASIS statements)	
	OWN	Sentences describing any aspect of the own work presented in this paper – except what is covered by AIM or TEXTUAL, e.g. details of solution (methodology), limitations, and further work.	
	AIM	Sentences best portraying the particular (main) research goal of the article	
	TEXTUAL	Explicit statements about the textual section structure of the paper	
	CONTRAST	Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons	
	BASIS	Statements that the own work uses some other work as its basis or starting point, or gets support from this other work	

Figure 1: Overview of the annotation scheme

1997; Alexandersson et al., 1995; Jurafsky et al., 1997), but our task is more difficult since it requires more subjective interpretation.

3 Annotation experiment

Our annotation scheme is based on the intuition that its categories provide an adequate and intuitive description of scientific texts. But this intuition alone is not enough of a justification: we believe that our claims, like claims about any other descriptive account of textual interpretation, should be substantiated by demonstrating that other humans can apply this interpretation consistently to actual texts.

We did three studies. Study I and II were designed to find out if the two versions of the annotation scheme (basic vs. full) can be learned by human coders with a significant amount of training. We are interested in two formal properties of the annotation scheme: stability and reproducibility (Krippendorff, 1980). Stability, the extent to which one annotator will produce the same classifications at different times, is important because an instable annotation scheme can never be reproducible. Reproducibility, the extent to which different annotators will produce the same classifications, is important because it measures the consistency of shared understandings (or meaning) held between annotators.

We use the Kappa coefficient K (Siegel and Castellan, 1988) to measure stability and reproducibility among k annotators on N items. In our experiment, the items are sentences. Kappa is a better measurement of agreement than raw percentage agreement (Carletta, 1996) because it factors out the level of agreement which would be reached by random annotators using the same distribution of categories as the real coders. 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. We expect high random agreement for our annotation scheme because so many sentences fall into the OWN category.

Studies I and II will determine how far we can trust in the human-annotated training material for both learning and evaluation of the automatic method. The outcome of Study II (full annotation scheme) is crucial to the task, as some of the categories specific to the full annotation scheme (particularly AIM) add considerable value to the information contained in the training material.

Study III tries to answer the question whether the considerable training effort used in Studies I and II can be reduced. If it were the case that coders with hardly any task-specific training can produce similar results to highly trained coders, the training material could be acquired in a more efficient way. A positive outcome of Study III would also strengthen claims about the intuitivity of the category definitions.

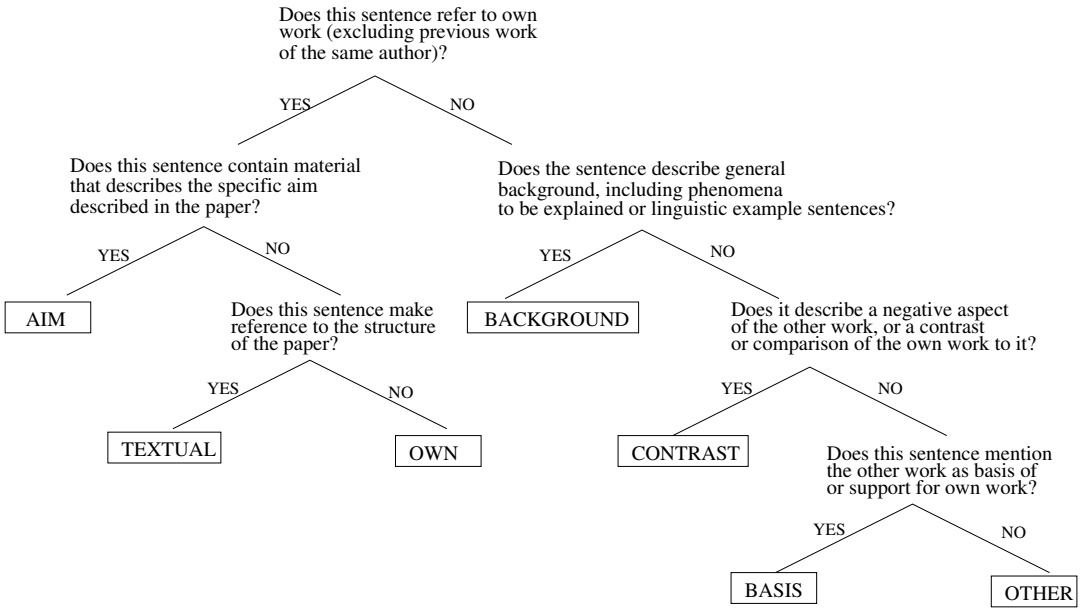

Figure 2: Decision tree for annotation

Our materials consist of 48 computational linguistics papers (22 for Study I, 26 for Study II), taken from the Computation and Language E-Print Archive (`http://xxx.lanl.gov/cmp-lg/`). We chose papers that had been presented at COLING, ANLP or ACL conferences (including student sessions), or ACL-sponsored workshops, and been put onto the archive between April 1994 and April 1995.

3.1 Studies I and II

For Studies I and II, we used three highly trained annotators. The annotators (two graduate students and the first author) can be considered skilled at extracting information from scientific papers but they were not experts in all of the subdomains of the papers they annotated. The annotators went through a substantial amount of training, including the reading of coding instructions for the two versions of the scheme (6 pages for the basic scheme and 17 pages for the full scheme), four training papers and weekly discussions, in which previous annotations were discussed. However, annotators were not allowed to change any previous decisions. For the stability figures (intra-annotator agreement), annotators re-coded 6 randomly chosen papers 6 weeks after the end of the annotation experiment. Skim-reading and annotation of an average length paper (3800 words) typically took the annotators 20–30 minutes.

During the annotation phase, one of the papers 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.

The results show that the basic annotation scheme is stable (K=.83, .79, .81; N=1248; k=2 for all three annotators) and reproducible (K=.78, N=4031, k=3). This reconfirms that trained annotators are capable of making the basic distinction between own work, specific other work, and general background. The full annotation scheme is stable (K=.82, .81, .76; N=1220; k=2 for all three annotators) and reproducible (K=.71, N=4261, k=3). Because of the increased cognitive difficulty of the task, the decrease in stability and reproducibility in comparison to Study I is acceptable. Leaving the coding developer out of the coder pool for Study II did not change the results (K=.71, N=4261, k=2), suggesting that the training conveyed her intentions fairly well.

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. As a general approach in our analysis, we wanted to look at the trends in the data as our main information source.

Figure 3 reports how well the four non-basic categories could be distinguished from all other categories, measured by Krippendorff's diagnostics for category distinctions (i.e. collapsing all *other* distinctions). When compared to the overall reproducibility of .71, we notice that the annotators were good at distinguishing AIM and TEX-

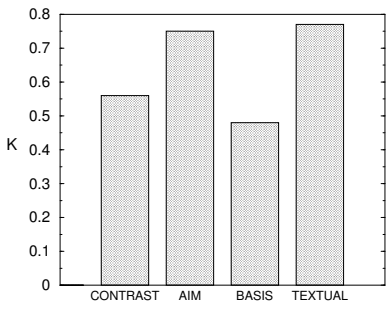

Figure 3: Reproducibility diagnostics: non-basic categories (Study II)

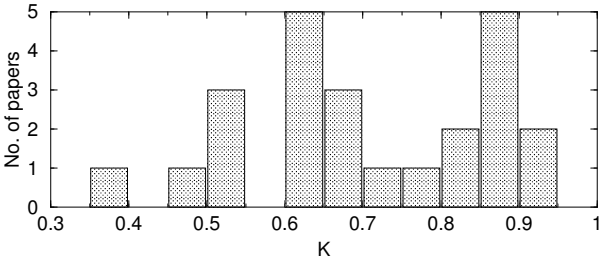

Figure 4: Distribution by reproducibility (Study II)

TUAL. This is an important result: as AIM sentences constitute the best characterization of the research paper for the summarization task we are particularly interested in having them annotated consistently in our training material. The annotators were less good at determining BASIS and CONTRAST. This might have to do with the location of those types of sentences in the paper: AIM and TEXTUAL are usually found at the beginning or end of the introduction section, whereas CONTRAST, and even more so BASIS, are usually interspersed within longer stretches of OWN. As a result, these categories are more exposed to lapses of attention during annotation.

If we blur the less important distinctions between CONTRAST, OTHER, and BACKGROUND, the reproducibility of the scheme increases to K=.75. Structuring our training set in this way seems to be a good compromise for our task, because with high reliability, it would still give us the crucial distinctions contained in the basic annotation scheme, plus the highly important AIM sentences, plus the useful TEXTUAL and BASIS sentences.

The variation in reproducibility across papers is large, both in Study I and Study II (cf. the quasi-bimodal distribution shown in Figure 4). Some hypotheses for why this might be so are the following:

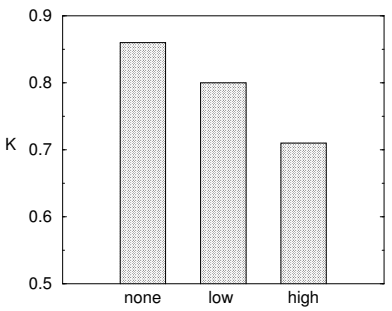

Figure 5: Effect of self-citation ratio on reproducibility (Study I)

- One 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 coded as OTHER) and *current, new* work. This was particularly the case where authors had published several papers about different aspects of one piece of research. We found a correlation with self citation ratio (ratio of self citations to all citations in running text): papers with many self citations are more difficult to annotate than papers that have few or no self citations (cf. Figure 5).

- Another persistent problematic distinction for our annotators was that between OWN and BACKGROUND. This could be a sign that some authors aimed their papers 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, it can indeed only be understood with a considerable amount of domain knowledge, which our annotators did not have.

- There is also a difference in reproducibility between papers from different *conference types*, as Figure 6 suggests. Out of our 25 papers, 4 were presented in student sessions, 4 came from workshops, the remaining 16 ones were main conference papers. Student session papers are easiest to annotate, which might be due to the fact that they are shorter and have a simpler structure, with less mentions of previous research. Main conference papers dedicate more space to describing and

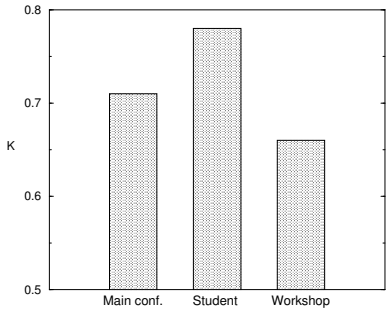

Figure 6: Effect of conference type on reproducibility (Study II)

criticising other people's work than student or workshop papers (on average about one fourth of the paper). They seem to be carefully prepared (and thus easy to annotate); conference authors must express themselves more clearly than workshop authors because they are reporting finished work to a wider audience.

3.2 Study III

For Study III, we used a different subject pool: 18 subjects with no prior annotation training. All of them had a graduate degree in Cognitive Science, with two exceptions: one was a graduate student in Sociology of Science, and one was a secretary. Subjects were given only minimal instructions (1 page A4), and the decision tree in Figure 2. Each annotator was randomly assigned to a group of six, all of whom independently annotated the same single paper. These three papers were randomly chosen from the set of papers for which our trained annotators had previously achieved good reproducibility in Study II (K=.65,N=205, k=3; K=.85,N=192,k=3; K=.87,N=144,k=3, respectively).

Reproducibility varied considerably between groups (K=.35, N=205, k=6; K=.49, N=192, k=6; K=.72, N=144, k=6). Kappa is designed to abstract over the number of coders. Lower reliablity for Study III as compared to Studies I and II is not an artefact of how K was calculated.

Some subjects in Group 1 and 2 did not understand the instructions as intended – we must conclude that our very short instructions did not provide enough information for consistent annotation. This is not surprising, given that human indexers (whose task is very similar to the task introduced here) are highly skilled professionals. However, part of this result can be attributed to the papers: Group 3, which annotated the paper found to be most reproducible in Study II, performed almost as well as trained annotators; Group 1, which performed worst, also happened to have the paper with the lowest reproducibility. In Groups 1 and 2, the most similar three annotators reached a respectable reproducibility (K=.5, N=205, k=3; K=.63, N=192, k=3). That, together with the good performance of Group 3, seems to show that the instructions did at least convey some of the meaning of the categories.

It is remarkable that the two subjects who had no training in computational linguistics performed reasonably well: they were not part of the circle of the three most similar subjects in their groups, but they were also not performing worse than the other two annotators.

4 Discussion

It is an interesting question how far shallow (human and automatic) information extraction methods, i.e. those using no domain knowledge, can be successful in a task such as ours. We believe that argumentative structure has so many reliable linguistic or non-linguistic correlates on the surface – physical layout being one of these correlates, others are linguistic indicators like *"to our knowledge"* and the relative order of the individual argumentative moves – that it should be possible to detect the line of argumentation of a text without much world knowledge. The two non-experts in the subject pool of Study III, who must have used some other information besides computational linguistics knowledge, performed satisfactorily – a fact that seems to confirm the promise of shallow methods.

Overall, reproducibility and stability for trained annotators does not quite reach the levels found for, for instance, the best dialogue act coding schemes (around K=.80). Our annotation requires more subjective judgments and is possibly more cognitively complex. Our reproducibility and stability results are in the range which Krippendorff (1980) describes as giving marginally significant results for reasonable size data sets when correlating two coded variables which would show a clear correlation if there were prefectly agreement. That is, the coding contains enough signal to be found among the noise of disagreement.

Of course, our requirements are rather less stringent than Krippendorff's because only one coded variable is involved, although coding 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. However, as with all coding schemes, its usefulness will only be clarified by the final appli-

cation.

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.

Most of the problems we identified in our studies have to do with a lack of distinction between own and other people's work (or own previous work). Because our scheme discriminates based on these properties, as well as being useful for summarizing research papers, it might be used for automatically detecting whether a paper is a review, a position paper, an evaluation paper or a 'pure' research article by looking at the relative frequencies of automatically annotated categories.

5 Conclusions

We have introduced an annotation scheme for research articles which marks the aims of the paper in relation to past literature. We have argued that this scheme is useful for building better abstracts, and have conducted some experiments which show that the annotation scheme can be learned by trained annotators and subsequently applied in a consistent way. Because the scheme is reliable, hand-annotated data can be used to train a system which applies the scheme automatically to unseen text.

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. As such, it should be independent of article length and article discipline. In the future, we plan to show this by applying our scheme to journal and conference articles from a range of disciplines. Practical reasons have kept us from using journal articles as data so far (namely the difficulty of corpus collection and the increased length and subsequent time effort of human experiments), but we are particularly interested in them as they can be expected to be of higher quality. As the basic argumentation is the same as in conference articles, our scheme should be applicable to journal articles at least as consistently as to the papers in our current collection.

6 Acknowledgements

We wish to thank our annotators, Vasilis Karaiskos and Ann Wilson, for their patience and diligence in this work, and for their insightful, critical, and very useful observations.

The first author is supported by an EPSRC studentship.

References

Jan Alexandersson, Elisabeth Maier, and Norbert Reithinger. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the Seventh European Meeting of the ACL*, pages 188–193.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Robin Cohen. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of COLING-84*, pages 251–255.

Anna Duszak. 1994. Academic discourse and intellectual styles. *Journal of Pragmatics*, 21:291–313.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca, 1997. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual.* University of Colorado, Institute of Cognitive Science. TR-97-02.

Joost G. Kircz. 1991. The rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation*, 47(4):354–372.

Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology.* Sage Commtext series; 5. Sage, Beverly Hills London.

Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference, Association for Computing Machinery, Special Interest Group Information Retrieval*, pages 68–73.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing and Management*, 27(1):55–81.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: description and construction of text structures. In G. Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 85–95, Dordrecht. Nijhoff.

Daniel Marcu. 1997. From discourse structures to text summaries. In Inderjeet Mani and Mark T. Maybury, editors, *Proceedings of the workshop on Intelligent Scalable Text Summarization, in association with ACL/EACL-97.*

Greg Myers. 1992. In this paper we report... – speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Chris D. Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26:171–186.

G.J Rath, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.

Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. 1989. Information extraction and text processing using linguistic knowledge acquisition. *Information Processing and Management*, 25(4):419–428.

Sidney Siegel and N.J. Jr. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences.* McGraw-Hill, second edition edition.

Karen Spärck Jones. 1998. Automatic summarising: factors and directions. In *ACL/EACL-97 Workshop 'Intelligent Scalable Text Summarization'.*

John Swales. 1990. *Genre analysis: English in academic and research settings.* Cambridge University Press.

Geoff Thompson and Yiyun Ye. 1991. Evaluation in the reporting verbs used in academic papers. *Applied Linguistics*, 12(4):365–382.

Sheryl R. Young and Phillip J. Hayes. 1985. Automatic classification and summarization of banking telexes. In *Proceedings of the Second Conference on Artificial Intelligence Applications.*