The Potsdam Commentary Corpus

Manfred Stede
University of Potsdam
Dept. of Linguistics
Applied Computational Linguistics
D-14415 Potsdam
Germany
stede@ling.uni-potsdam.de

Abstract

A corpus of German newspaper commentaries has been assembled and annotated with different information (and currently, to different degrees): part-of-speech, syntax, rhetorical structure, connectives, co-reference, and information structure. The paper explains the design decisions taken in the annotations, and describes a number of applications using this corpus with its multi-layer annotation.

1 Introduction

A corpus of German newspaper commentaries has been assembled at Potsdam University, and annotated with different linguistic information, to different degrees. Two aspects of the corpus have been presented in previous papers ((Reitter, Stede 2003) on underspecified rhetorical structure; (Stede 2003) on the perspective of knowledge-based summarization). This paper, however, provides a comprehensive overview of the data collection effort and its current state.

At present, the 'Potsdam Commentary Corpus' (henceforth 'PCC' for short) consists of 170 commentaries from *Märkische Allgemeine Zeitung*, a German regional daily. The choice of the genre *commentary* resulted from the fact that an investigation of *rhetorical structure*, its interaction with other aspects of discourse structure, and the prospects for its automatic derivation are the key motivations for building up the corpus. Commentaries argue in favor of a specific point of view toward some political issue, often dicussing yet dismissing other points of view; therefore, they typically offer a more interesting rhetorical structure than, say, narrative text or other portions of newspapers.

The choice of the particular newspaper was motivated by the fact that the language used in a regional daily is somewhat simpler than that of papers read nationwide. (Again, the goal of automatic analysis was responsible for this decision.) This is manifest in the lexical choices but also in structural features. As an indication, in our core corpus, we found an average sentence length of 15.8 words and 1.8 verbs per sentence, whereas a randomly taken sample of ten commentaries from the national papers *Süddeutsche Zeitung* and *Frankfurter Allgemeine* has 19.6 words and 2.1 verbs per sentence. The commentaries in PCC are all of roughly the same length, ranging from 8 to 10 sentences. For illustration, an English translation of one of the commentaries is given in Figure 1.

The paper is organized as follows: Section 2 explains the different layers of annotation that have been produced or are being produced. Section 3 discusses the applications that have been completed with PCC, or are under way, or are planned for the future. Section 4 draws some conclusions from the present state of the effort.

2 Layers of Annotation

The corpus has been annotated with six different types of information, which are characterized in the following subsections. Not all the layers have been produced for all the texts yet. There is a 'core corpus' of ten commentaries, for which the range of information (except for syntax) has been completed; the remaining data has been annotated to different degrees, as explained below.

All annotations are done with specific tools and in XML; each layer has its own DTD. This offers the well-known advantages for interchangability, but it raises the question of how to query the corpus across levels of annotation. We will briefly discuss this point in Section 3.1.

2.1 Part-of-speech tags

All commentaries have been tagged with part-of-speech information using Brants' *TnT*[1] tagger and the *Stuttgart/Tübingen Tag Set*

[1] www.coli.uni-sb.de/~thorsten/tnt/

Dagmar Ziegler is up to her neck in debt. Due to the dramatic fiscal situation in Brandenburg she now surprisingly withdrew legislation drafted more than a year ago, and suggested to decide on it not before 2003. Unexpectedly, because the ministries of treasury and education both had prepared the teacher plan together. This withdrawal by the treasury secretary is understandable, though. It is difficult to motivate these days why one ministry should be exempt from cutbacks — at the expense of the others. Reiche's colleagues will make sure that the concept is waterproof. Indeed there are several open issues. For one thing, it is not clear who is to receive settlements or what should happen in case not enough teachers accept the offer of early retirement. Nonetheless there is no alternative to Reiche's plan. The state in future has not enough work for its many teachers. And time is short. The significant drop in number of pupils will begin in the fall of 2003. The government has to make a decision, and do it quickly. Either save money at any cost - or give priority to education.

Figure 1: Translation of PCC sample commentary

(STTS)[2].

2.2 Syntactic structure

Annotation of syntactic structure for the core corpus has just begun. We follow the guidelines developed in the TIGER project (Brants et al. 2002) for syntactic annotation of German newspaper text, using the *Annotate*[3] tool for interactive construction of tree structures.

2.3 Rhetorical structure

All commentaries have been annotated with rhetorical structure, using *RSTTool*[4] and the definitions of discourse relations provided by Rhetorical Structure Theory (Mann, Thompson 1988). Two annotators received training with the RST definitions and started the process with a first set of 10 texts, the results of which were intensively discussed and revised. Then, the remaining texts were annotated and cross-validated, always with discussions among the annotators. Thus we opted not to take the step of creating more precise written annotation guidelines (as (Carlson, Marcu 2001) did for English), which would then allow for measuring inter-annotator agreement. The motivation for our more informal approach was the intuition that there are so many open problems in rhetorical analysis (and more so for German than for English; see below) that the main task is qualitative investigation, whereas rigorous quantitative analyses should be performed at a later stage.

One conclusion drawn from this annotation effort was that for humans and machines alike, assigning rhetorical relations is a process loaded with ambiguity and, possibly, subjectivity. We respond to this on the one hand with a format for its *underspecification* (see 2.4) and on the other hand with an additional level of annotation that attends only to connectives and their scopes (see 2.5), which is intended as an intermediate step on the long road towards a systematic and objective treatment of rhetorical structure.

2.4 Underspecified rhetorical structure

While RST (Mann, Thompson 1988) proposed that a single relation hold between adjacent text segments, SDRT (Asher, Lascarides 2003) maintains that multiple relations may hold simultaneously. Within the RST "user community" there has also been discussion whether two levels of discourse structure should not be systematically distinguished (intentional versus informational).

Some relations are signalled by subordinating conjunctions, which clearly demarcate the range of the text spans related (matrix clause, embedded clause). When the signal is a coordinating conjunction, the second span is usually the clause following the conjunction; the first span is often the clause preceding it, but sometimes stretches further back. When the connective is an adverbial, there is much less clarity as to the range of the spans.

Assigning rhetorical relations thus poses questions that can often be answered only subjectively. Our annotators pointed out that very often they made almost random decisions as to what relation to choose, and where to locate the boundary of a span. (Carlson, Marcu 2001) responded to this situation with relatively precise (and therefore long!) annotation guidelines that tell annotators what to do in case of doubt.

[2] www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

[3] www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

[4] www.wagsoft.com/RSTTool

Quite often, though, these directives fulfill the goal of increasing annotator agreement without in fact settling the theoretical question; i.e., the directives are clear but not always very well motivated.

In (Reitter, Stede 2003) we went a different way and suggested *URML*[5], an XML format for *underspecifying* rhetorical structure: a number of relations can be assigned instead of a single one, competing analyses can be represented with shared forests. The rhetorical structure annotations of PCC have all been converted to URML. There are still some open issues to be resolved with the format, but it represents a first step. What ought to be developed now is an annotation tool that can make use of the format, allow for underspecified annotations and visualize them accordingly.

2.5 Connectives with scopes

For the 'core' portion of PCC, we found that on average, 35% of the coherence relations in our RST annotations are explicitly signalled by a lexical connective.[6] When adding the fact that connectives are often ambiguous, one has to conclude that prospects for an automatic analysis of rhetorical structure using shallow methods (i.e., relying largely on connectives) are not bright — but see Sections 3.2 and 3.3 below. Still, for both human and automatic rhetorical analysis, connectives are the most important source of surface information. We thus decided to pay specific attention to them and introduce an annotation layer for connectives and their scopes. This was also inspired by the work on the Penn Discourse Tree Bank[7], which follows similar goals for English.

For effectively annotating connectives/scopes, we found that existing annotation tools were not well-suited, for two reasons:

- Some tools are dedicated to modes of annotation (e.g., tiers), which could only quite un-intuitively be used for connectives and scopes.
- Some tools would allow for the desired annotation mode, but are so complicated (they can be used for many other purposes as well) that annotators take a long time getting used to them.

Consequently, we implemented our own annotation tool *ConAno* in Java (Stede, Heintze 2004), which provides specifically the functionality needed for our purpose. It reads a file with a list of German connectives, and when a text is opened for annotation, it highlights all the words that show up in this list; these will be all the potential connectives. The annotator can then "click away" those words that are here not used as connectives (such as the conjunction *und* ('and') used in lists, or many adverbials that are ambiguous between connective and discourse particle). Then, moving from connective to connective, *ConAno* sometimes offers suggestions for its scope (using heuristics like 'for subjunctor, mark all words up to the next comma as the first segment'), which the annotator can accept with a mouseclick or overwrite, marking instead the correct scope with the mouse. When finished, the whole material is written into an XML-structured annotation file.

2.6 Co-reference

We developed a first version of annotation guidelines for co-reference in PCC (Gross 2003), which served as basis for annotating the core corpus but have not been empirically evaluated for inter-annotator agreement yet. The tool we use is *MMAX*[8], which has been specifically designed for marking co-reference.

Upon identifying an anaphoric expression (currently restricted to: pronouns, prepositional adverbs, definite noun phrases), the annotator first marks the antecedent expression (currently restricted to: various kinds of noun phrases, prepositional phrases, verb phrases, sentences) and then establishes the link between the two. Links can be of two different kinds: anaphoric or bridging (definite noun phrases picking up an antecedent via world-knowledge).

- Anaphoric links: the annotator is asked to specify whether the anaphor is a repetition, partial repetition, pronoun, epithet (e.g., *Andy Warhol – the PopArt artist*), or is-a (e.g., *Andy Warhol was often hunted by photographers. This fact annoyed especially his dog...*).
- Bridging links: the annotator is asked to specify the type as part-whole, cause-effect (e.g., *She had an accident. The wounds are still healing.*), entity-attribute (e.g., *She*

[5] 'Underspecified Rhetorical Markup Language'

[6] This confirms the figure given by (Schauer, Hahn 2001), who determined that in their corpus of German computer tests, 38% of relations were lexically signalled.

[7] www.cis.upenn.edu/∼pdtb/

[8] www.eml-research.de/english/Research/NLP/Downloads

had to buy a new car. The price shocked her.), or same-kind (e.g., *Her health insurance paid for the hospital fees, but the automobile insurance did not cover the repair.*).

2.7 Information structure

In a similar effort, (Götze 2003) developed a proposal for the theory-neutral annotation of information structure (IS) — a notoriously difficult area with plenty of conflicting and overlapping terminological conceptions. And indeed, converging on annotation guidelines is even more difficult than it is with co-reference. Like in the co-reference annotation, Götze's proposal has been applied by two annotators to the core corpus but it has not been systematically evaluated yet.

We use *MMAX* for this annotation as well. Here, annotation proceeds in two phases: first, the *domains* and the *units* of IS are marked as such. The domains are the linguistic spans that are to receive an IS-partitioning, and the units are the (smaller) spans that can play a role as a constituent of such a partitioning. Among the IS-units, the referring expressions are marked as such and will in the second phase receive a label for *cognitive status* (active, accessible-text, accessible-situation, inferrable, inactive). They are also labelled for their topicality (yes / no), and this annotation is accompanied by a confidence value assigned by the annotator (since it is a more subjective matter). Finally, the focus/background partition is annotated, together with the focus question that elicits the corresponding answer. Asking the annotator to also formulate the question is a way of arriving at more reproducible decisions.

For all these annotation taks, Götze developed a series of questions (essentially a decision tree) designed to lead the annotator to the appropriate judgement.

3 Past, Present, Future Applications

Having explained the various layers of annotation in PCC, we now turn to the question what all this might be good for. This concerns on the one hand the basic question of retrieval, i.e. searching for information across the annotation layers (see 3.1). On the other hand, we are interested in the application of rhetorical analysis or 'discourse parsing' (3.2 and 3.3), in text generation (3.4), and in exploiting the corpus for the development of improved models of discourse structure (3.5).

3.1 Retrieval

For displaying and querying the annoated text, we make use of the *Annis* Linguistic Database developed in our group for a large research effort ('Sonderforschungsbereich') revolving around information structure.[9] The implementation is basically complete, yet some improvements and extensions are still under way. The web-based *Annis* imports data in a variety of XML formats and tagsets and displays it in a tier-oriented way (optionally, trees can be drawn more elegantly in a separate window). Figure 2 shows a screenshot (which is of somewhat limited value, though, as color plays a major role in signalling the different statuses of the information). In the small window on the left, search queries can be entered, here one for an NP that has been annotated on the co-reference layer as *bridging*. The portions of information in the large window can be individually clicked visible or invisible; here we have chosen to see (from top to bottom)

- the full text,
- the annotation values for the activated annotation set (co-reference),
- the actual annotation tiers, and
- the portion of text currently 'in focus' (which also appears underlined in the full text).

Different annotations of the same text are mapped into the same data structure, so that search queries can be formulated across annotation levels. Thus it is possible, for illustration, to look for a noun phrase (syntax tier) marked as topic (information structure tier) that is in a bridging relation (co-reference tier) to some other noun phrase.

3.2 Stochastic rhetorical analysis

In an experiment on automatic rhetorical parsing, the RST-annotations and PoS tags were used by (Reitter 2003) as a training corpus for statistical classification with Support Vector Machines. Since 170 annotated texts constitute a fairly small training set, Reitter found that an overall recognition accuracy of 39% could be achieved using his method. For the English RST-annotated corpus that is made available via LDC, his corresponding result is 62%. Future work along these lines will incorporate other layers of annotation, in particular the syntax information.

[9] www.ling.uni-potsdam.de/sfb/

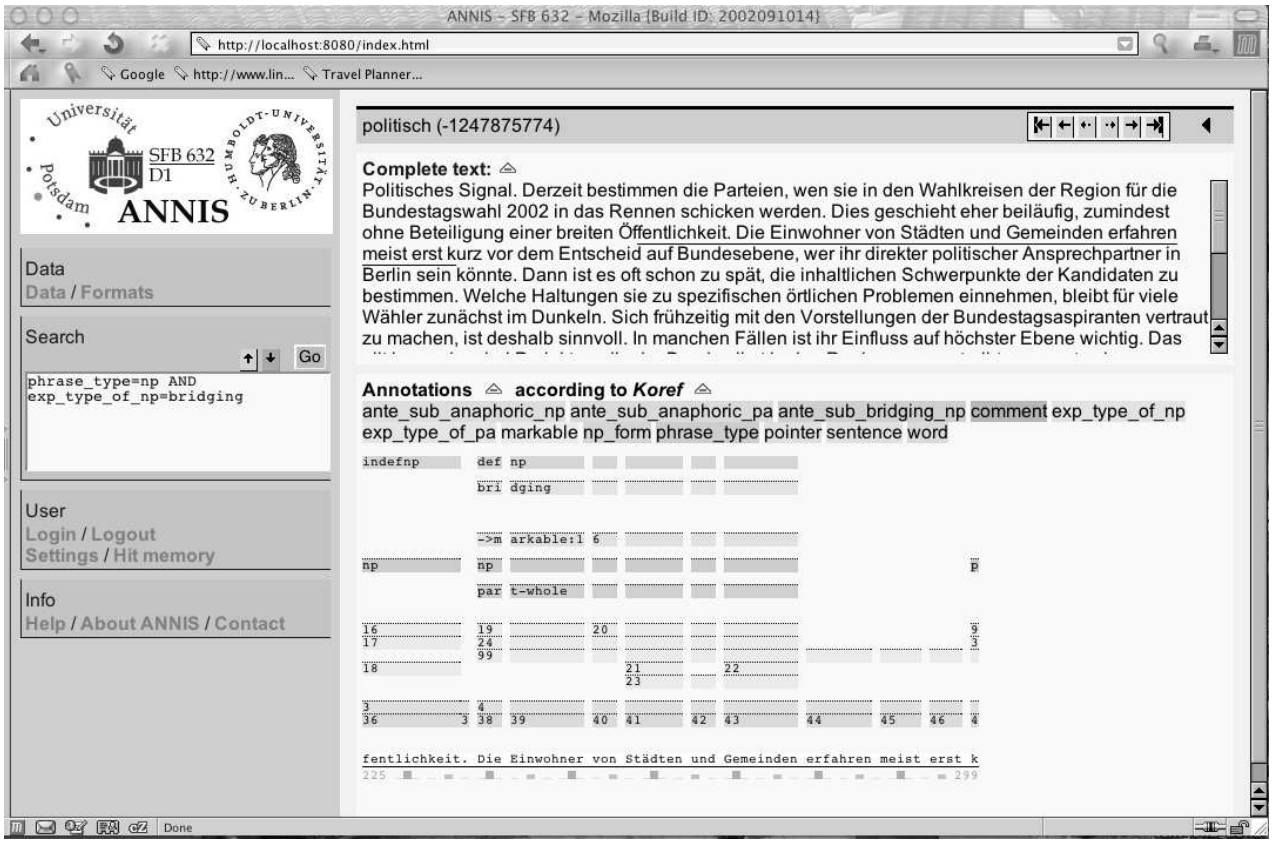

Figure 2: Screenshot of *Annis* Linguistic Database

3.3 Symbolic and knowledge-based rhetorical analysis

We are experimenting with a hybrid statistical and knowledge-based system for discourse parsing and summarization (Stede 2003), (Hanneforth et al. 2003), again targeting the genre of commentaries. The idea is to have a pipeline of shallow-analysis modules (tagging, chunking, discourse parsing based on connectives) and map the resulting underspecified rhetorical tree (see Section 2.4) into a knowledge base that may contain domain and world knowledge for enriching the representation, e.g., to resolve references that cannot be handled by shallow methods, or to hypothesize coherence relations. In the rhetorical tree, nuclearity information is then used to extract a "kernel tree" that supposedly represents the key information from which the summary can be generated (which in turn may involve co-reference information, as we want to avoid dangling pronouns in a summary). Thus we are interested not in extraction, but actual generation from representations that may be developed to different degrees of granularity.

In order to evaluate and advance this approach, it helps to feed into the knowledge base data that is already enriched with some of the desired information — as in PCC. That is, we can use the discourse parser on PCC texts, emulating for instance a "co-reference oracle" that adds the information from our co-reference annotations. The knowledge base then can be tested for its relation-inference capabilities on the basis of full-blown co-reference information. Conversely, we can use the full rhetorical tree from the annotations and tune the co-reference module. The general idea for the knowledge-based part is to have the system use as much information as it can find at its disposal to produce a target representation as specific as possible and as underspecified as necessary. For developing these mechanisms, the possibility to feed in hand-annotated information is very useful.

3.4 Salience-based text generation

Text generation, or at least the two phases of text planning and sentence planning, is a process driven partly by well-motivated choices (e.g., use this lexeme X rather than that more colloquial near-synonym Y) and partly by conventionalized patterns (e.g., order of information in news reports). And then there are decisions that systems typically hard-wire, because the linguistic motivation for making them is not well understood yet. Preferences for constituent order (especially in languages with relatively free word order) often belong to this group. Trying to integrate constituent ordering and choice of referring expressions, (Chiarcos 2003) developed a numerical model of *salience propagation* that captures various factors of author's intentions and of information structure for ordering sentences as well as smaller constituents, and picking appropriate referring expressions.[10] Chiarcos used the PCC annotations of co-reference and information structure to compute his numerical models for salience projection across the generated texts.

3.5 Improved models of discourse structure

Besides the applications just sketched, the overarching goal of developing the PCC is to build up an empirical basis for investigating phenomena of discourse structure. One key issue here is to seek a discourse-based model of information structure. Since Daneš' proposals of 'thematic development patterns', a few suggestions have been made as to the existence of a level of discourse structure that would predict the information structure of sentences within texts. (Hartmann 1984), for example, used the term *Reliefgebung* to characterize the distibution of main and minor information in texts (similar to the notion of nuclearity in RST). (Brandt 1996) extended these ideas toward a conception of *kommunikative Gewichtung* ('communicative-weight assignment'). A different notion of information structure, is used in work such as that of (**?**), who tried to characterize felicitous constituent ordering (theme choice, in particular) that leads to texts presenting information in a natural, "flowing" way rather than with abrupt shifts of attention. — In order to ground such approaches in linguistic observation and description, a multi-level annotation like that of PCC can be exploited to look for correlations in particular between syntactic structure, choice of referring expressions, and sentence-internal information structure.

A different but supplementary perspective on discourse-based information structure is taken by one of our partner projects[11], which is interested in correlations between prosody and discourse structure. A number of PCC commentaries will be read by professional news speakers and prosodic features be annotated, so that the various annotation layers can be set into correspondence with intonation patterns. In focus is in particular the correlation with rhetorical structure, i.e., the question whether specific rhetorical relations — or groups of relations in particular configurations — are signalled by speakers with prosodic means.

Besides information structure, the second main goal is to enhance current models of rhetorical structure. As already pointed out in Section 2.4, current theories diverge not only on the number and definition of relations but also on apects of structure, i.e., whether a tree is sufficient as a representational device or general graphs are required (and if so, whether any restrictions can be placed on these graph's structures — cf. (Webber et al., 2003)). Again, the idea is that having a picture of syntax, co-reference, and sentence-internal information structure at one's disposal should aid in finding models of discourse structure that are more explanatory and can be empirically supported.

4 Conclusions

The PCC is not the result of a funded project. Instead, the designs of the various annotation layers and the actual annotation work are results of a series of diploma theses, of students' work in course projects, and to some extent of paid assistentships. This means that the PCC cannot grow particularly quickly. After the first step towards breadth had been taken with the PoS-tagging, RST annotation, and URML conversion of the entire corpus of 170 texts[12], emphasis shifted towards depth. Hence we decided to select ten commentaries to form a 'core corpus', for which the entire range of annotation levels was realized, so that experiments with multi-level querying could commence. Cur-

[10] For an exposition of the idea as applied to the task of text planning, see (Chiarcos, Stede 2004).

[11] www.ling.uni-potsdam.de/sfb/projekt_a3.php

[12] This step was carried out in the course of the diploma thesis work of David Reitter (2003), which deserves special mention here.

rently, some annotations (in particular the connectives and scopes) have already moved beyond the core corpus; the others will grow step by step.

The kind of annotation work presented here would clearly benefit from the emergence of standard formats and tag sets, which could lead to sharable resources of larger size. Clearly this poses a number of research challenges, though, such as the applicability of tag sets across different languages. Nonetheless, the prospect of a network of annotated discourse resources seems particularly promising if not only a single annotation layer is used but a whole variety of them, so that a systematic search for correlations between them becomes possible, which in turn can lead to more explanatory models of discourse structure.

References

N. Asher, A. Lascarides. 2003. *Logics of Conversation.* Cambridge University Press.

M. Brandt. 1996. Subordination und Parenthese als Mittel der Informationsstrukturierung in Texten. In. W. Motsch (ed.): *Ebenen der Textstruktur.* Tübingen: Niemeyer.

S. Brants, S. Dipper, S. Hansen, W. Lezius, G. Smith. 2002. The TIGER Treebank. In: Proc. of the Workshop on Treebanks and Linguistic Theories. Sozopol.

L. Carlson, D. Marcu. 2001. Discourse Tagging Reference Manual. Ms., Univ. of Southern California / Information Sciences Institute.

C. Chiarcos. 2003. Eine Satzplanungskomponente für die Textgenerierung. In: Uta Seewald-Heeg (ed.): *Sprachtechnologie für die multilinguale Kommunikation.* Bonn: gardez. (Short version of Diploma Thesis, Technische Universität Berlin)

C. Chiarcos, M. Stede. 2004. Salience-Driven Text Planning. To appear in: Proc. of the Third Int'l Conference on Natural Language Generation (INLG), Careys Manor (UK).

P. Fries. 1981. On the Status of Theme in English. *Forum Linguisticum* 6.1:1-38.

M. Götze. 2003. Zur Annotation von Informationsstruktur. Diploma thesis, Universität Potsdam, Inst. of Linguistics.

J. Gross. 2003. Algorithmen zur Behandlung von Anaphora in Zeitungskommentaren. Diploma thesis, Technische Universität Berlin.

T. Hanneforth, S. Heintze, M. Stede. 2003. Rhetorical Parsing with Underspecification and Forests. In: Proc. of the HLT/NAACL Conference (Companion Volume), Edmonton/AL.

D. Hartmann. 1984. Reliefgebung: Informationsvordergrund und Informationshintergrund in Texten als Problem von Textlinguistik und Stilistik. In: *Wirkendes Wort* 34:305-323.

W. Mann, S. Thompson. 1988. Rhetorical Structure Theory: A Theory of Text Organization. *TEXT* 8(3):243-281.

D. Reitter. 2003. Simple signals for complex rhetorics: on rhetorical analysis with rich-feature support vector modelds. In: Uta Seewald-Heeg (ed.): *Sprachtechnologie für die multilinguale Kommunikation.* Bonn: gardez. (Short version of Diploma Thesis, Universität Potsdam, Inst. of Linguistics)

D. Reitter, M. Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In: Proc. of the 4th Int'l Workshop on Linguistically Interpreted Corpora (LINC-03), Budapest.

H. Schauer, U. Hahn. 2001. Anaphoric cues for coherence relations. In: Proc. of 'Recent Advances in Natural Language Processing'-RANLP 2001. Tzigov Chark, Bulgaria.

M. Stede. 2003. Surfaces and depths in text understanding: the case of newspaper commentary. In: Proc. of the HLT/NAACL Workshop on Text Meaning, Edmonton/AL.

M. Stede, S. Heintze. 2004. Machine-assisted rhetorical structure annotation. To appear in: Proc. of the 20th Int'l Conference on Computational Linguistics (COLING), Geneva.

B. Webber, A. Knott, M. Stone, A. Joshi. 2003. Anaphora and Discourse Structure. *Computational Linguistics* 29(4):545-588.