Surfaces and Depths in Text Understanding: The Case of Newspaper Commentary

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

Abstract

Using a specific example of a newspaper commentary, the paper explores the relationship between 'surface-oriented' and 'deep' analysis for purposes such as text summarization. The discussion is followed by a description of our ongoing work on automatic commentary understanding and the current state of the implementation.

1 Introduction

Generally speaking, language understanding for some cognitive agent means reconstructing the presumed speaker's goals in communicating with her/him/it. An application-specific automatic system might very well hard-wire some or most of the aspects of this reconstruction process, but things get more interesting when the complexity is acknowledged and paid attention to. When moving from individual utterances to understanding connected discourse, an additional problem arises: that of partitioning the material into segments (usually at various levels) and that of inferring the connections between text segments (or between their underlying illocutions).

In recent years, some surface-based approaches to "rhetorical parsing" have been proposed, which try to recover a text's discourse structure, following the general layout of Rhetorical Structure Theory (Mann, Thompson, 1988). Starting from this idea, in this paper, we imagine to push the goal of rhetorical parsing a bit further. The idea is that of a system that can take a newspaper commentary and understand it to the effect that it can, amongst other things, produce the "most concise summary" of it:

- the topic of the commentary
- the position the author is taking toward it

This goal does not seem reachable with methods of shallow analysis alone. But why exactly is it not, and what methods are needed in addition? In the following, we work through a sample commentary and analyse the steps and the knowledge necessary to arrive at the desired result, i.e., a concise summary. Thereafter, we sketch the state of our implementation work, which follows the goal of fusing surface-based methods with knowledge-based analysis.

2 Sample commentary

Figure 1 shows a sample newspaper commentary, taken from the German regional daily "Märkische Allgemeine Zeitung" in October 2002, along with an English translation. To ease reference, numbers have been inserted in front of the sentences. Let us first move through the text and make some clarifications so that the reader can get the picture of what is going on. Dagmar Ziegler is the treasury secretary of the German state of Brandenburg. A plan for early retirement of teachers had been drafted collectively by her and the education secretary, whose name is Reiche. Sentence 5 points out that the plan had intended education to be exempt from the cutbacks happening all over the various ministries — *Reiche's colleagues* in 6 are thus the other secretaries. While the middle part of the text provides some motivation for the withdrawal, 9-14 state that the plan nonetheless should be implemented, for the reasons given in 10-12. Our intended "most concise summary" then would be:

- Topic: Treasury secretary delays decision on teacher staff plan
- Author's opinion: Government has to decide quickly and give priority to education, thus implement the plan

Notice that a statistical summarization technique (i.e., a sentence extraction approach) is very unlikely to yield

(1) Dagmar Ziegler sitzt in der Schuldenfalle. (2) Auf Grund der dramatischen Kassenlage in Brandenburg hat sie jetzt eine seit mehr als einem Jahr erarbeitete Kabinettsvorlage überraschend auf Eis gelegt und vorgeschlagen, erst 2003 darüber zu entscheiden. (3) Überraschend, weil das Finanz- und das Bildungsressort das Lehrerpersonalkonzept gemeinsam entwickelt hatten. (4) Der Rückzieher der Finanzministerin ist aber verständlich. (5) Es dürfte derzeit schwer zu vermitteln sein, weshalb ein Ressort pauschal von künftigen Einsparungen ausgenommen werden soll - auf Kosten der anderen. (6) Reiches Ministerkollegen werden mit Argusaugen darüber wachen, dass das Konzept wasserdicht ist. (7) Tatsächlich gibt es noch etliche offene Fragen. (8) So ist etwa unklar, wer Abfindungen erhalten soll, oder was passiert, wenn zu wenig Lehrer die Angebote des vorzeitigen Ausstiegs nutzen. (9) Dennoch gibt es zu Reiches Personalpapier eigentlich keine Alternative. (10) Das Land hat künftig zu wenig Arbeit für zu viele Pädagogen. (11) Und die Zeit drängt. (12) Der große Einbruch der Schülerzahlen an den weiterführenden Schulen beginnt bereits im Herbst 2003. (13) Die Regierung muss sich entscheiden, und zwar schnell. (14) Entweder sparen um jeden Preis - oder Priorität fuer die Bildung.

(1) Dagmar Ziegler is up to her neck in debt. (2) 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. (3) Unexpectedly, because the ministries of treasury and education both had prepared the teacher plan together. (4) This withdrawal by the treasury secretary is understandable, though. (5) It is difficult to motivate these days why one ministry should be exempt from cutbacks — at the expense of the others. (6) Reiche's colleagues will make sure that the concept is waterproof. (7) Indeed there are several open issues. (8) 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. (9) Nonetheless there is no alternative to Reiche's plan. (10) The state in future has not enough work for its many teachers. (11) And time is short. (12) The significant drop in number of pupils will begin in the fall of 2003. (13) The government has to make a decision, and do it quickly. (14) Either save money at any cost - or give priority to education.

Figure 1: Sample text with translation

a result along these lines, because word frequency is of little help in cases where the line of the argument has to be pulled out of the text, and might make some synthesis necessary. Just to illustrate the point, the Microsoft Word "25 percent" summarization reads as follows:

> *Überraschend, weil das Finanz- und das Bildungsressort das Lehrerpersonalkonzept gemeinsam entwickelt hatten. Reiches Ministerkollegen werden mit Argusaugen darüber wachen, dass das Konzept wasserdicht ist. Entweder sparen um jeden Preis - oder Priorität für die Bildung.*
>
> *Unexpectedly, because the ministries of treasury and education both had prepared the teacher plan together. Reiche's colleagues will make sure that the concept is waterproof. Either save money at any cost - or give priority to education.*

It includes the final sentence (most probably because it *is* the final sentence), but in the context of the other two extracted sentences it does not convey the author's position — nor the precise problem under discussion.

3 Rhetorical Structure

Since RST (Mann, Thompson 1988) has been so influential in discourse-oriented computational linguistics, we start our analysis with a "man-made" RST analysis, which was produced collectively by two RST-experienced students. See Figure 2.[1] (The English reader can relatively easy map the German segments to their translations in Fig. 1 with the help of the sentence numbers added to the text in the tree).

Some considerations motivating this analysis (in terms of segment numbers, *not* sentence numbers): 1 is seen as the general Background for the satellite of the overall Concession, which discusses the problem arising from the debt situation. Arguably, it might as well be treated as Background to the entire text. The Evaluation between 2-6 and 7-12 is a relation often found in opinion texts; an alternative to be considered here is Antithesis — in this case, however, 7-12 would have to be the nucleus, which seems to be problematic in light of the situation that 3-4 is the main portion that is being related to the material in 13-16.

8-12 explains and elaborates the author's opinion that the withdrawal is understandable (7). The distinctions between the relations Explanation, Elaboration, and Evidence were mostly based on surface cues, such as *tatsächlich* ('indeed') signalling Evidence. The Elabora-

[1] Visualization by the RST Tool (O'Donnell, 1997). Notation follows Mann and Thompson (1988): vertical bars and incoming arrows denote nuclear segments, outgoing arrows denote satellites. Numbers at leaves are sentence numbers; segment numbers are given at internal nodes.

tions, on the other hand, take up one aspect from the previous utterance and provide additional information, such as the two open questions in 10-12.

13-16 then overwrites this acknowledged "understanding" of Ziegler's move and states that her plan should be implemented anyway, and that the matter is urgent. It is here where the kernel of the author's opinion on the matter is located (and argued for by 14-16). The final part 17-20 then is a little less decisive, re-states the urgency, and uses a 'rhetorical alternative' in 19-20 to indirectly indicate that the plan should be implemented, education be given priority.

Rhetorical analysis is anything but an uncontroversial matter. For our purposes, though, let us take the proposed analysis as the point of departure for subsequent considerations. We first have to ask whether such an RST tree is indeed significant and useful for the goals of text understanding as outlined in Section 1 — and should this question receive an affirmative answer, we need to turn to the prospects for automating the analysis.

4 The role of RST trees in text understanding

Does the information encoded in Figure 2 make a contribution to our needs? Yes, fortunately it does. First of all, investigating the lengths of the lines beginning from the top, we notice that the RST tree contains a useful segmentation of the text. Its main constituents are segments 1, 2-6, 7-12, 13-16, and 17-20. Next, we are given a set of central nuclei coming from these constituents: 3/4, 7, 13, and 17. Finally, we find the most obvious ingredient of an RST analysis: coherence relations. When we proceed to extract the relations that connect our main constituents and then replace each constituent with (a paraphrase of) its central nucleus, we are left with the RST tree shown in Figure 3. This tree, assuming that it also determines the linear order of the text units, can be verbalized in English for instance like this:

> *That Ziegler withdrew the legislation on teacher staff is understandable; nonetheless, there is no alternative to it. The Brandenburg government must make a decision now.*

This, it seems, is not bad for a concise summary of the text. Notice furthermore that additional material from the original tree can be added to the extracted tree when desired, such as the reason for act A being understandable (incrementally segments 8, 9, 10, 11-12).

We initially conclude that a rhetorical tree seems to be useful as a backbone for a text representation, based on which we can perform operations such as summarization. While we are not the first to point this out (see, e.g., Marcu 1999), we shall now move on to ask how one

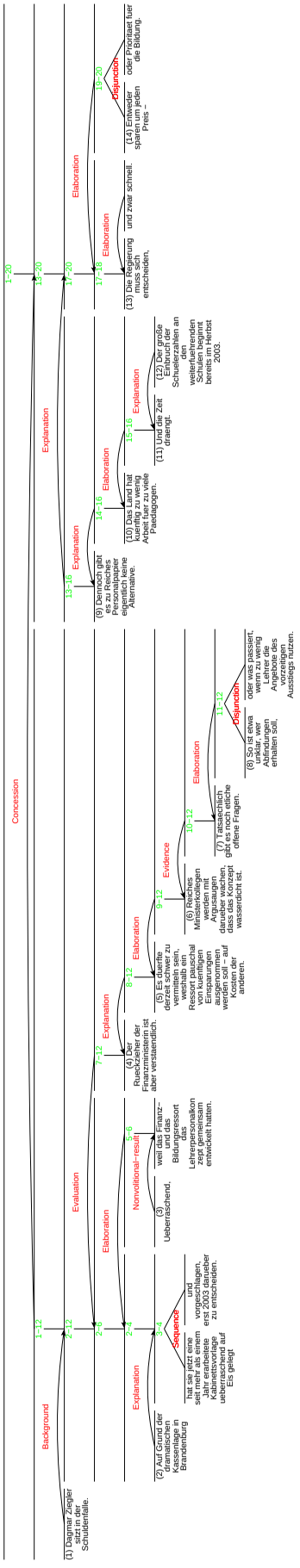

Figure 2: RST tree for sample text

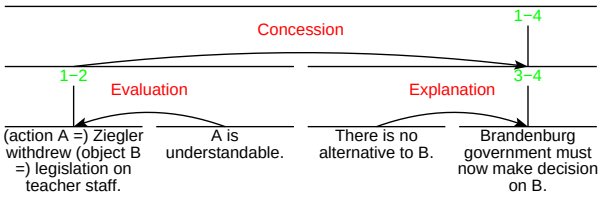

Figure 3: Desired "summary tree" for sample text

would arrive at such a tree — more specifically, at a formal representation of it.

What kind of information is necessary beyond assigning relations, spans and nuclei? In our representation of the summary tree, we have implicitly assumed that *reference resolution* has been worked out - in particular that the legislation can be identified in the satellite of the Explanation, and also in its nucleus, where it figures implicitly as the object to be decided upon. Further, an RST tree does not explicitly represent the *topic* of the discourse, as we had asked for in the beginning. In our present example, things happen to work out quite well, but in general, an explicit topic identification step will be needed. And finally, the rhetorical tree does not have information on illocution types (1-place rhetorical relations, so to speak) that distinguish reported facts (e.g., segments 3 and 4) from author's opinion (e.g., segment 7). We will return to these issues in Section 6, but first consider the chances for building up rhetorical trees automatically.

5 Prospects for Rhetorical Parsing

Major proponents of rhetorical parsing have been (Sumita et al., 1992), (Corston-Oliver, 1998), (Marcu, 1997), and (Schilder, 2002). All these approaches emphasise their membership in the "shallow analysis" family; they are based solely on surface cues, none tries to work with semantic / domain / world knowledge. (Corston-Oliver and Schilder use some genre-specific heuristics for preferential parsing, though.) In general, our sample text belongs to a rather "friendly" genre for rhetorical parsing, as commentaries are relatively rich in connectives, which are the most important source of information for making decisions — but not the only one: Corston-Oliver, for example, points out that certain linguistic features such as modality can sometimes help disambiguating connectives. Let us now hypothesize what an "ideal" surface-oriented rhetorical parser, equipped with a good lexicon of connectives, part-of-speech tagger and some rough rules of phrase composition, could do with our example text.

5.1 Segmentation

As we are imagining an "ideal" shallow analyser, it might very well produce the segmentation that is underlying the human analysis in Figure 2. The obvious first step is to establish a segment boundary at every full stop that terminates a sentence (no ambiguities in our text). Within sentences, there are six additional segment boundaries, which can be identified by considering connectives and part-of-speech tags of surrounding words, i.e. by a variant of "chunk parsing": *Auf Grund* ('due to') has to be followed by an NP and establishes a segment up to the finite verb (*hat*). The *und* ('and') can be identified to conjoin complete verb phrases and thus should trigger a boundary. In the following sentence, *weil* ('because') has to be followed by a full clause, forming a segment. The next intra-sentential break is between segments 11 and 12; the *oder* ('or') can be identified like the *und* above. In segment 17-18, *und zwar* ('and in particular') is a strict boundary marker, as is the *entweder – oder* ('either – or') construction in 19-20.

5.2 Relations, scopes, nuclei

The lexical boundary markers just mentioned also indicate (classes of) rhetorical relationships. *Auf Grund* — when used in its idiomatic reading — signals some kind of Cause with the satellite following in an NP. Because the *und* in 3-4 co-occurs with the temporal expressions *jetzt* ('now') and *erst 2003* ('not before 2003'), it can be taken as a signal of Sequence here, with the boundaries clearly identifiable, so that the RST subtree 2-4 can be derived fully. Furthermore, 5 takes up a single adverbial *überraschend* from 3, and in conjunction with the *weil*-clause in 6, the Elaboration can be inferred. *weil* ('because') itself signals some Cause, but the nuclearity decision (which in the "real" tree in Fig. 2 leads to choosing Result) is difficult here; since 5 merely repeats a lexeme from 3, we might assign nuclearity status to 6 on the "surface" grounds that it is longer and provides new material. We thus have derived a rhetorical structure for the entire span 2-6. In 7, *aber* ('but') should be expected to signal either Contrast or Concession; how far the leftmost span reaches can not be determined, though. Both 8 and 9 provide no reliable surface clues. In 10, *tatsächlich* ('indeed') can be taken as an adverbial indicating Evidence; again the scope towards the left is not clear. *So .. etwa* ('thus .. for instance') in 11 marks an Elaboration, and the *oder* in 12 a Disjunction between the two clauses. Span 10-12 therefore receives an analysis. In 13, *dennoch* ('nonetheless') is a clear Concession signal, but its scope cannot be reliably determined. Finally, the only two remaining decisions to be made from surface observations are the Elaboration 17-18 (*und zwar*, 'and in particular') and the Disjunction 19-20. Then, making use of RST's "empty" relation Join, we can bind together the assembled pieces and are left with the tree shown in Fig. 4.

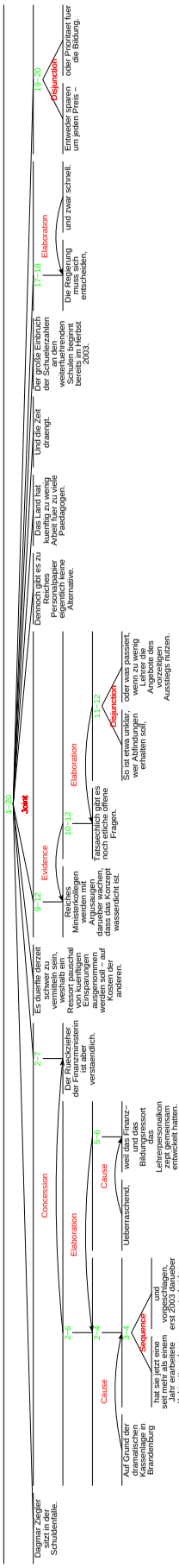

Figure 4: Result of “surface parsing” of sample text

5.3 Heuristics or statistics

In the analysis just proposed, we used lexical knowledge (connectives – relations) as well as some linguistic cues. In addition, rhetorical parsers can either apply domain- or genre-specific heuristics, or hypothesize further relations by employing probabilistic knowledge gathered from training with annotated corpora. What can be expected to be gained in this way for our sample text?

Since the unanalysed 1 is followed by a larger segment, we might hypothesize 1 to be a Background for following material; this is certainly common in commentaries. The satellite of Contrast/Concession to the left of 7 can be assumed to be the larger segment preceding it; how far the nucleus stretches to the right is difficult to see, though. Statistically, it will likely be only segment 8. The situation is similar with the Concession hypothesized at 13 — it is somewhat likely (though wrong in this case!) that the nucleus will be only the segment hosting the connective, but about the satellite span nothing can be said here. Finally, at the very end of the commentary, a heuristic might tell that it should not terminate with a binuclear disjunction as a prominent nucleus (such a commentary would probably fail to make a point), and hence it seems advisable to treat 19-20 as a satellite of a larger span 17-20, and a “defensive” relation guess would be Elaboration.

Returning to the issue of segmentation, we can also try to apply surface-based heuristic methods to finding larger segments, i.e., to split the text into its major parts, which has sometimes been called “text tiling”. For instance, a boundary between “macro segments” 13-16 and 17-20 is hinted at by the definite NP *Die Regierung* (’the government’) at the beginning of 17, which has no antecedent NP in the preceding segment and hence can be interpreted as a change of discourse topic. Such considerations can be unreliable, though. *Schuldenfalle* (’up to the neck in debt’) and *dramatische Kassenlage* (‘dramatic fiscal situation’) seem to bind 1 and 2 closely together, and yet there is a major segment boundary in our tree in Fig. 2.

5.4 Assessment

Under the assumption that our discussion reasonably reflects the state of the art in surface-oriented analysis methods, we now have to compare its result to our overall target, the summary tree in Figure 3. We have successfully found segment 3-4 as the central nucleus of the span 2-6, and we have hypothesized it being related to 7 (without identifying the Evaluation relation). As for the other half of the target tree, 17 has been hypothesized as an important nucleus, but we have no clear connection to 13 (its target satellite), as the “staircase” of Elaborations and Explanations 13-16 could not be identified. Nor could we determine the central role of the Concession that combines the key nuclei.

At this point, we can draw three intermediate conclu-

sions. First, rhetorical parsing should allow for *underspecified* representations as — intermediate or final — outcome; see (Hanneforth et al., submitted). Second, text understanding aiming at quality needs to go further than surface-oriented rhetorical parsing. With the help of additional domain/world-knowledge sources, attempts should be made to fill gaps in the analysis. It is then an implementation decision whether to fuse these additional processes into the rhetorical parser, or to use a pipeline approach where the parser produces an underspecified rhetorical tree that can afterwards be further enriched. Third, probabilistic or statistical knowledge can also serve to fill gaps, but the information drawn from such sources should be marked with its status being insecure. As opposed to decisions based on lexical/linguistic knowledge (in 5.2), the tentative decisions from 5.3 may be overwritten by later knowledge-based processes.

6 Knowledge-Based Understanding

"Understanding a text" for some cognitive agent means to fuse prior knowledge with information encountered in the text. This process has ramifications for both sides: What I know or believe influences what exactly it is that I "take away" from a text, and my knowledge and beliefs will usually to a certain extent be affected by what I read. Naturally, the process varies from agent to agent: They will understand different portions of a text in different ways and to different degrees. Thus, when we endeavour to devise and implement models of text understanding, the target should not be to arrive at "the one and only" result, but rather to account for the mechanics of this variability: the *mechanism* of understanding should be the same, but the *result* depend on the type and amount of prior knowledge that the agent carries. In the end, a representation of text meaning should therefore be designed to allow for this flexibility.

6.1 KB Design

In line with many approaches to using knoweldge for language processing, we adopt the framework of *terminological logic* as the vehicle for representing both the background knowledge necessary to bootstrap any understanding process, and the content of the text. Thus the basic idea is to encode prior, general knowledge in the TBox (concepts) and the information from the text in the ABox (instances). For our example, the subworld of government, ministries and legislation has to be modelled in the TBox, so that entities referred to in the text can instantiate the appropriate concepts. We thus map the rhetorical tree built up by shallow analysis to an ABox in the LOOM language (MacGregor, Bates, 1987); for a sketch of representing rhetorical structure in LOOM, see (Stede, 1999, ch. 10).

6.2 "Ideal" text understanding

Each leaf of the tree is now subject to detailled semantic analysis and mapped to an enriched predicate/argument structure that instantiates the relevant portions of the TBox (quite similar to the 'Text Meaning Representation' of (Mahesh, Nirenburg, 1996)). "Enriched" indicates that beyond the plain proposition, we need information such as modality but also the type of illocution; e.g., does the utterance represent a factual statement, the author's opinion, or a proposal? This is necessary for analyzing the structure of an argument (but, of course, often it is very difficult to determine).

One central task in text understanding is reference resolution. Surface-based methods can perform initial work here, but without some background knowledge, the task can generally not be completed. In our sample text, understanding the argument depends on recognizing that *Kabinettsvorlage* in (2), *Lehrerpersonalkonzept* in (3), *Konzept* in (6), and *Reiches Personalpapier* in (9) all refer to the same entity; that *Ziegler* in (1) and *Finanzministerin* in (4) are co-referent; that *Finanz- und Bildungsressort* in (3), *Reiches Ministerkollegen* in (6), and *die Regierung* in (13) refer to portions of or the complete Brandenburg government, respectively. Once again, hints can be derived from the surface words (e.g., by compund analysis of *Lehrerpersonalkonzept*), but only background knowledge (an ontology) about the composition of governments and their tasks enables the final decisions.

Knowledge-based inferences are necessary to infer rhetorical relations such as Explanation or Evaluation. Consider for example segment 15-16, where the relationship between 'time is short' (a subjective, evaluative statement) and 'begin already in the fall of 2003' (a statement of a fact), once recognized, prompts us to assign Explanation. Similarly, the Elaboration between this segment and the preceeding 14 can be based on the fact that 14 makes a statement about the 'future situation' in Brandenburg, which is made more specific by time being short and the fall of 2003. More complex inferences are necessary to attach 14-16 then to 13 (and similarly in the segment 7-12).

6.3 "Realistic" text understanding

Even if it were possible to hand-code the knowledge base such that for our present sample text the complete representation can be constructed — for the general text analysis situation, achieving a performance anywhere near the "complete and correct solution" is beyond reach. As indicated at the beginning of the section, though, this is not necessarily bad news, as a notion of partial understanding, or "mixed-depth encoding" as suggested by Hirst and Ryan (1992), should be the rule rather than the exception. Under ideal circumstances, a clause at a leaf of the rhetorical tree might be fully analyzed, with all refer-

ences resolved and no gaps remaining. In the worst case, however, understanding might fail entirely. Then, following Hirst and Ryan, the text portion itself should simply be part of the representation. In most cases, the representation will be somewhere in-between: some aspects fully analyzed, but others not or incompletely understood. For example, a sentence adverbial might be unknown and thus the modality of the sentence not be determined. The ABox then should reflect this partiality accordingly, and allow for appropriate inferences on the different levels of representation.

The notion of mixed depth is relevant not only for the tree's leaves: Sometimes, it might not be possible to derive a unique rhetorical relation between two segments, in which case a set of candidates can be given, or none at all, or just an assignment of nucleus and satellite segments, if there are cues allowing to infer this. In (Reitter and Stede, 2003) we suggest an XML-based format for representing such underspecified rhetorical structures.

Projecting this onto the terminological logic scheme, and adding the treatment of leaves, we need to provide the TBox not only with concepts representing entities of "the world" but also with those representing linguistic objects, such as *clause* or *noun group*, and for the case of unanalyzed material, *string*. To briefly elaborate the *noun group* example, consider *Reiches Ministerkollegen* ('Reiche's colleagues') in sentence 6. Shallow analysis will identify *Reiche* as some proper name and thus the two words as a noun group. An ABox istance of this type is created, and it depends on the knowledge held by the TBox whether additional types can be inferred. Reiche has not been mentioned before in the text, because from the perspective auf the author the name is prominent enough to be identified promptly by the (local) readers. If the system's TBox contains a person of that name in the domain of the Brandenburg government, the link can be made; otherwise, *Reiche* will be some un-identified object about which the ABox collects some information from the text.

Representations containing material with different degrees of analysis become useful when accompanied by *processes* that are able to work with them ('mixed-depth processing'). For summarization, this means that the task becomes one of fusing *extraction* (of unanalyzed portions that have been identified as important nuclei) with *generation* (from the representations of analyzed portions). Of course, this can lead to errors such as dangling anaphors in the extracted portions, but that is the price we pay for robustness — robustness in this refined sense of "analyse as deeply as you can" instead of the more common "extract *something* rather than fail."

7 Implementation Strategy

Finally, here is a brief sketch of the implementation work that is under way in the Computational Linguistics group at Potsdam University. Newspaper commentaries are the genre of choice for most of our current work. We have assembled a corpus of some 150 commentaries from "Märkische Allgemeine Zeitung", annotated with rhetorical relations, using the RST Tool by O'Donnell (1997). It uses an XML format that we convert to our format of underspecified rhetorical structure ('URML' Reitter & Stede 2003).

This data, along with suitable retrieval tools, informs our implementation work on automatic commentary understanding and generation. Focusing here on understanding, our first prototype (Hanneforth et al., submitted) uses a pipeline of modules performing

1. tokenization
2. sentence splitting and segmentation into clauses
3. part-of-speech tagging
4. chunk parsing
5. rhetorical parsing
6. knowledge-based processing

The tagger we are using is the Tree Tagger by Schmid (1994); the chunk parser is CASS (Abney 1996). The remaining modules, as well as the grammars for the chunk parser, have been developed by our group (including student projects).[2] The rhetorical parser is a chart parser and uses a discourse grammar leading to a parse forest, and is supported by a lexicon of discourse markers (connectives). We have started work on reference resolution (in conjunction with named-entity recognition). Addition of the knowledge-based component, as sketched in the previous section, has just begun. The main challenge is to allow for the various kinds of underspecification within the LOOM formalism and to design appropriate inference rules.

As implementation shell, we are using GATE (http://www.gate.ac.uk), which proved to be a very useful environment for this kind of incremental system construction.

8 Conclusions

Knowledge-based text understanding and surface-based analysis have in the past largely been perceived as very different enterprises that do not even share the same

[2] In addition to this "traditional" pipeline approach, Reitter (2003) performed experiments with machine learning techniques based on our MAZ corpus as training data.

goals. The paper argued that a synthesis can be useful, in particular: that knowledge-based understanding can benefit from stages of surface-based pre-processing. Given that

- pre-coded knowledge will almost certainly have gaps when it comes to understanding a "new" text, and
- surface-based methods yield "some" analysis for any text, however sparse, irrelevant or even wrong that analysis may be,

a better notion of *robustness* is needed that explains how language understanding can be "as good (deep) as possible or as necessary". The proposal is to first employ "defensive" surface-based methods to provide a first, underspecified representation of text structure that has gaps but is relatively trustworthy. Then, this representation may be enriched with the help of statistical, probabilistic, heuristic information that is added to the representation (and marked as being less trustworthy). Finally, a "deep" analysis can map everything into a TBox/ABox scheme, possibly again filling some gaps in the text representation (Abox) on the basis of prior knowledge already encoded in the TBox. The deep analysis should not be an all-or-nothing step but perform as good as possible — if something cannot be understood entirely, then be content with a partial representation or, in the worst case, with a portion of the surface string.

Acknowledgements

Thanks to: Thomas Hanneforth and all the students of our *Systemkonstruktion* seminar for the implementation of the rhetorical parser prototype; anonymous reviewers for helpful comments on the paper; *Märkische Allgemeine Zeitung* for providing us with plenty of commentaries.

References

Abney, S. 1996. Partial Parsing via Finite-State Cascades. In: Proceedings of the ESSLLI '96 Robust Parsing Workshop.

Corston-Oliver, S. 1998. Computing representations of the structure of written discourse. Ph.D. Thesis. University of California, Santa Barbara.

Hanneforth, T.; Heintze, S.; Stede, M. Rhetorical parsing with underspecification and forests. Submitted.

Hirst, G.; Ryan, M. 1992. Mixed-depth representations for natural language text. In: P. Jacobs (ed.): *Text-based intelligent systems*. Lawrence Erlbaum, Hillsdale.

MacGregor, R.; Bates, R. 1987. The LOOM Knowledge Representation Language. Technical Report ISI/RS-87-188, USC Information Sciences Institute.

Mahesh, K.; Nirenburg, S.; 1996. Meaning representation for knowledge sharing in practical machine translation. Proc. of the FLAIRS-96 track on information interchange; Florida AI Research Symposium, Key West.

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

Marcu, D. 1997. The rhetorical parsing of natural language texts. Proc. of the 35th Annual Conference of the ACL, 96-103.

Marcu, D. 1999. Discourse trees are good indicators of importance in text. In: I. Mani and M. Maybury (eds.): *Advances in Automatic Text Summarization*, 123-136, The MIT Press.

O'Donnell, M. 1997. RST-Tool: An RST Analysis Tool. Proc. of the 6th European Workshop on Natural Language Generation, Duisburg.

Reitter, D. 2003. Rhetorical analysis with rich-feature support vector models. Diploma Thesis, Potsdam University, Dept. of Linguistics.

Reitter, D.; Stede, M. 2003. Step by step: underspecified markup in incremental rhetorical analysis In: Proc. of the Worksop on Linguistically Interpreted Corpora (LINC-03), Budapest.

Schilder, F. 2002. Robust Discourse Parsing via Discourse Markers, Topicality and Position. *Natural Language Engineering* 8 (2/3).

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. Proc. of the Int'l Conference on New Methods in Language Processing.

Stede, M. 1999. *Lexical Semantics and Knowledge Representation in Multilingual Text Generation*. Kluwer, Dordrecht/Boston.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. 1992. A discourse structure analyzer for Japanese text. Proc. of the International Conference on Fifth Generation Computer Systems, 1133-1140.