Originally published in: Butt, Miriam (Hrsg.): Proceedings of KONVENS 2006
(Konferenz zur Verarbeitung natürlicher Sprache). - Konstanz: Universität Konstanz, 2006. Pp. 52-55.

A Text-technological Approach to Automatic Discourse Analysis of Complex Texts

Mirco Hilbert **Henning Lobin** **Maja Bärenfänger**

Harald Lüngen **Csilla Puskás**

Institut für Germanistik
Arbeitsbereich Angewandte Sprachwissenschaft und Computerlinguistik
Justus-Liebig-Universität Gießen

Abstract

This paper describes the development of a relational discourse parsing architecture for text documents of a complex text type, namely scientific articles. To achieve this goal, several different linguistic knowledge sources and auxiliary analyses on different linguistic levels are necessary.

Introduction

Automatic discourse analysis is an open research field where various different approaches exist. Often, these approaches deal with uncomplex, less structured texts such as newspaper articles deriving their discourse structure as an application of a relational discourse theory. Here, the analysis of discourse connectives like conjunctions and adverbs in addition to morphological and syntactic features of the input texts are used as bases for the relational discourse parsing.

When analysing documents of more complex text types, additional sources of knowledge are needed to detect the relational discourse structure of the texts. In our ongoing research project we are developing a discourse parser using the logical document structure, the genre-specific text type structure and the thematic structure as abstract discourse markers in addition to the traditional discourse connectives and morpho-syntactic features. The input of our parser are scientific articles as a complex text type. Therefore, we built a corpus of 120 scientific articles from two different disciplines (psychology and linguistics), two different languages (English and German) and two different subgenres (experimental and review). We assumed that all three factors have their influences in the different text structures and therefore also in the derivation of the relational discourse structure.

In Section 1, this paper first describes the linguistic foundations of the different relevant text structures, which we are analysing and processing in our project. Proceeding from these theoretical foundations, in Section 2 the architecture of our discourse parser is introduced with its concrete analysis components and the resulting discourse structure representation format.

1 Linguistic foundations

1.1 Logical document structure

Documents can be regarded as complex segments which are hierarchically built up from smaller segments. In a syntagmatic perspective documents are described by grammars containing rules which define the way in which segments can be combined to yield valid documents of a certain type. Moreover, the logical document structure is interrelated with the graphical layout structure of a document. According to the document grammar, complex segments can be formed by a compositional aggregation of adjacent segments.

1.2 Discourse structure

The discourse structure of a text can be represented by a system of discourse coherence relations which hold between the individual segments of the text. In a relational discourse theory these text segments are called "spans" and include elementary discourse segments (EDSs) as well as complex discourse segments (CDSs), which are relationally structured themselves.

Several text type-independent discourse theories exist to represent and built up these discourse structures. These are for example the Unified Linguistic Discourse Model ULDM (Polanyi et al., 2004a; Polanyi et al., 2004b), the Segmented Discourse Representation Theory SDRT (Asher and Lascarides, 2003; Asher and Vieu, 2005), and the Rhetorical Structure Theory RST (Mann and Thompson, 1988; Marcu, 2000). In each theory the discourse structure is represented by a hierarchy of subordinating or coordinating relations. But it is not commonly agreed whether the structure should be a graph or a tree. In SDRT the discourse structure is represented by a graph, in ULDM and RST a tree structure is used. In our system we developed a tree-like discourse representation format based on RST, which, though, can contain cross-references in some clearly defined cases. The specific format is further described in Section 2.2.

1.2.1 Set of rhetorical relations

While in ULDM the discourse tree only consists of the two types of relations, subordination and coordination, in RST the nature of these relations can be further described by labels of differently defined rhetorical relations. In RST a subordinating relation is called "mononuclear" (or "hypotactic") because it consists of exactly one nucleus preceded or followed by one satellite (Marcu, 2000; Corston-Oliver, 1998; Egg and Redeker, 2005). According to Carlson et al. (2001) the nucleus of a rhetorical relation is "the more salient, essential piece of information". Coordinating relations are called "multinuclear" (or "paratactic") consisting of two or more nuclei. 26 text type- and application-independent rhetorical relations were introduced by Mann and Thompson (1988) whereas they are at the same time open to be adjusted for special text types or applications. Later on, they have been grouped and classified by Hovy and Maier (1995) and Carlson et al. (2001) to construct taxonomies of rhetorical relations; see also (Goecke et al., 2005).

Investigations of our corpus of scientific articles and manual test annotations have shown that some of these

relations are not relevant for our text type and some others were missing or could be subclassified for our annotation purposes. The resulting set of rhetorical relations is hierarchically structured where each super-relation is an abstraction of its sub-relations and can be used instead by means of ambiguity. Also their relation types are defined, whether they are mononuclear, multinuclear or bi-nuclear, i. e. multinuclear with exactly two nuclei. The set consists of 27 ideational, 22 interpersonal, and 13 other rhetorical relations. These can be reduced to a set of 22 higher level relations.

1.2.2 Discourse marker lexicon

As with the different discourse theories, there exist different approaches of discourse analyses and the construction of these formal discourse representations. In SDRT the fully-fledged semantic representation of a discourse segment is needed in a logical form for being combined with other discourse segments by logical aggregation methods. In RST the discourse segments (spans) are plain text, whereas the aggregation of spans presupposes knowledge about the meaning of these spans. As a complete semantic interpretation is difficult in the computational analysis of discourse structures, other information is used to anticipate the relation of two spans. This necessary information is often obtained by auxiliary text analyses deriving linguistic properties such as discourse connectives and morpho-syntactic features (Corston-Oliver, 1998; Marcu, 2000; Polanyi et al., 2004a).

To analyse the text of a scientific article and recognize these clues of rhetorical relations, we manually developed (and further are extending) a set of lexical and also abstract discourse markers which are indicating possible rhetorical relations. Each discourse marker entry consists of a cue, such as a pattern how the discourse marker can be detected when appearing in a text, an optional filter, which defines additional obligatory conditions that militate in favour of a rhetorical relation, and a set of one or more rhetorical relations that possibly come along with that discourse marker. Abstract discourse markers or abstract ancillary conditions of lexical discourse markers are morpho-syntactic features that can be obtained by an automatical grammatical tagger (see Section 2.3), and also document-logical, thematic, and text type-specific features as discussed in the following sections.

1.3 Thematic structure

The thematic structure is one component of two levels of discourse structure. Using RST, it is possible to analyse and represent both levels, the local level – by annotating the relations between sequential elementary discourse segments – as well as the global level – by relating complex discourse segments. Especially for these last-mentioned relations across larger spans of text, the ideational relation Elaboration is particularly useful (Carlson et al., 2001).

To perform these thematic structure analyses we will use the lexico-semantic net GermaNet (Kunze and Lemnitzer, 2002) as one knowledge source, which indicates semantic relations between concepts (see also Section 2.3). These can be used as cues for the recognition of the ELABORATION-RELATION. For a detailed modelling of thematic relations an extension of the ELABORATION-RELATION with different subtypes is necessary. Based on an analysis of our corpus we identified a set of different Elaboration subtypes, which are relevant for our discourse annotation tasks.

The relation ELABORATION-IDENTITY holds between a nucleus and a satellite that share a referential identity, that are about the same discourse referent. On the one hand we distinguish between forms of theme-theme or rheme-theme chaining (cf. Polanyi et al. (2003)), on the other hand between ASSIGNMENT (of a technical term or an abbreviation) and other forms of specification, where the meaning of the theme in the nucleus is expanded, restricted or specified by its satellite. ELABORATION-SPECIFICATION is similar to the relation ELABORATION-OBJECT-ATTRIBUTE as used by Carlson and Marcu (2001).

ELABORATION-DERIVATION comprises all relations between a nucleus and a satellite which are based on topic derivation, composition, ontological subordination or coordination. The subtypes of this relation are all mentioned in various publications but have never been grouped together (Mann and Thompson, 1988; Hovy and Maier, 1995; Carlson and Marcu, 2001).

1.4 Text type structure

The genre-specific text type structure is a major clue for the global argumentative and rhetorical structure of a complex text. Discourse relations between text spans on higher levels of discourse, such as section-combining relations, are not usually signalled by lexical or grammatical discourse markers. Text type-specific topic types as described by van Dijk (1980), Swales (1990) and Teufel (1999) include functional categories such as "introduction", "background", "method", "result" and "conclusion".

These topic types act as functions between the assigned parts of the text and the text type of the whole text. Therefore, the topic type of a text span can be used as a clue for a rhetorical relation between this and another text span. For instance a text span classified as "conclusion" may be a satellite of a rhetorical relation "Result" within the discourse structure of the text. Thus, a text type structure analysis, as already been realized with automatic text categorisation methods (Kando, 1999; Teufel and Moens, 2002; Langer et al., 2004), is a significant supplier of relational cues on a macro level of discourse analysis.

To analyse and annotate the text type structure we developed a hierarchically organised text type structure schema of 135 topic types categories and, additionally, automatically derived a reduced schema of 21 categories, which is more suitable for an efficient and consistent annotation.

2 Discourse parser architecture

The architecture of the relational discourse parser, which we are developing in our research project, is shown in Figure 1. Its realization can be divided in three major version steps. In each version the complexity of the parser architecture will be increased and it takes more knowledge sources as input to guide its decisions. For the representation and processing of the several knowledge sources and the text linguistic analyses on the different linguistic levels, text-technological XML-based formalisms and methods are employed.

The declarative knowledge sources are used in several preprocessing steps by auxiliary analysis components to

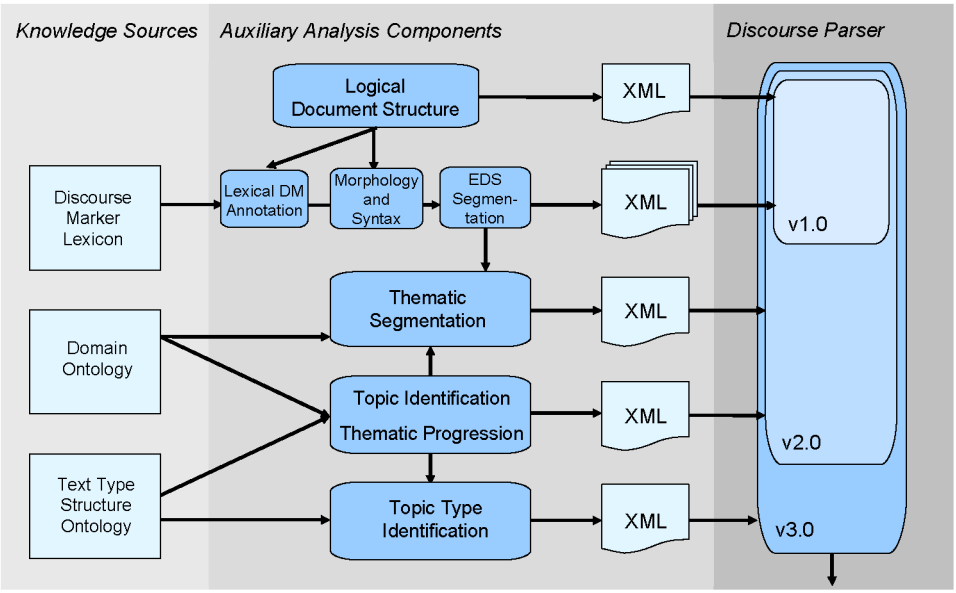

Figure 1: Discourse parser architecture.

analyse and annotate the input text. The discourse parser itself takes these different annotations as its input and may access the knowledge sources again according to its parsing strategy. The output of the parser will be a set of possible relational discourse structures of the input text as discussed in sections 1.2 and 2.2.

In the following, the different knowledge sources, the resulting output format of the parser and its architectural structure consisting of the several analysis and parsing components are discussed in more detail.

2.1 Knowledge Sources

For the version 1.0 of the discourse parser two main resources are used as the knowledge sources for the auxiliary analysis components and the parser itself: the taxonomy of rhetorical relations (called "RRSet"), which we described in Section 1.2.1, and the discourse marker lexicon of lexical, morpho-syntactic and abstract discourse markers, which indicate hypotheses of rhetorical relations (called "DMList"), described in Section 1.2.2. For both sources purpose-built XML representation formats have been developed. The taxonomy of elaboration relations described in Section 1.3 is part of the set of rhetorical relations and therefore also included in the RRSet.

For versions 2.0 and 3.0 of the parser the text type structure schema described in Section 1.4 is also provided as XML. Additionally, a domain ontology of discipline-specific terms is to be developed as an addition or extension of the lexico-semantic net for German GermaNet (Kunze and Lemnitzer, 2002) and the linguistic terminological ontology GOLD (Farrar and Langendoen, 2003).

2.2 Discourse structure representation

To represent the resulting relational discourse structure of our parser an XML format has been developed. It is called the HP format where HP stands for the two central relation types: hypotactic (mononuclear) and paratactic (multinuclear) rhetorical relations. In contrast to the Underspecified Rhetorical Markup Language URML (Reitter and Stede, 2003), where the basic tree structure of the discourse analysis is accomplished by XML ID referencing techniques, the HP format uses the given XML document structure to represent the discourse tree. Thus, the HP format is not just a data format to store the discourse structure information but a text-technological annotation of the underlying input text.

Since scientific articles are long and highly structured natural language texts, there are some peculiarities, where the basic tree structure is insufficient. These comprise "dislocated satellites", like footnotes and floating objects such as figures and tables, which are not adjacent to their nuclei, "embedded satellites", like parentheses or other rhetorically relevant insertions, and "multiple dependencies", where secondary but relevant rhetorical relations overlap with the relations of the basic discourse structure. To handle these clearly defined extensions to the main tree structure, special elements and XML ID referencing techniques are introduced.

2.3 Discourse parser components

The component to analyse the "Logical Document Structure" takes a scientific article as input and generates a logical document structure annotation encoded in an extended subset of the DocBook format (Walsh and Muellner, 1999). The "Lexical DM Annotation" component uses the knowledge of the DMList to annotate discourse markers occurring in the input text. The "Morphology and Syntax" component uses the grammatical tagger Machinese Syntax by Connexor Oy, based on the Functional Dependency Grammar by Tapanainen and Järvinen (1997), to analyse and annotate morphological and sentence-syntactic features. The output of this tool is converted to a customized XML representation required by our parsing system. Finally, the "EDS Segmentation" component is a discourse segmenter, which segments the input text into elementary, sentence and complex discourse segments using the discourse marker lexicon, punctuation and morpho-syntactic features as knowledge sources to detect the boundings of the discourse segments.

For versions 2.0 and 3.0 the parser needs three additional auxiliary analysis components. The "Topic Identification and Thematic Progression" component analy-

ses the text according to thematic and text type clues (see sections 1.3 and 1.4). The results of these analyses are also used for the thematic structure analysis in the "Thematic Segmentation" component and for the text type structure analysis in the "Topic Type Identification" component.

All analysis components output a corresponding XML annotation of the input text. This set of primary data identical annotations yields a multi-layered XML representation (cf. Witt et al. (2005)), which is the input of the discourse parser. The discourse parser itself uses a bootstrapping technique to iteratively generate new or revised XML annotation layers based on the available input annotation layers. The final output format of the parser will be a set of possible rhetorical discourse structures in the HP format described in the previous Section 2.2. For the parsing process a shift-reduced algorithm is used, which is based on an approach by Marcu (2000).

3 Summary

In this paper we described the linguistic foundations and our resulting development and design decisions in the process of realizing a relational discourse parsing system for complex texts. We described the problems which arose from analysing the different linguistic levels of scientific articles and possible solutions. We are currently developing the version 1.0 of the parsing system.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation.* Cambridge University Press, Cambridge, UK.

Nicholas Asher and Laure Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*, Eurospeech 2001, Denmark.

Simon Corston-Oliver. 1998. *Computing of Representations of the Structure of Written Discourse.* Ph.D. thesis, University of California, Santa Barbara.

Markus Egg and Gisela Redeker. 2005. Underspecified discourse representation. In Claudia Sassen, Anton Benz, and Peter Kühnlein, editors, *Proceedings of Constraints in Discourse*, pages 46–53, Dortmund.

Scott Farrar and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

Daniela Goecke, Harald Lüngen, Felix Sasaki, Andreas Witt, and Scott Farrar. 2005. GOLD and discourse: Domain- and community-specific extensions. In *Proceedings of the 2005 E-MELD-Workshop*, Boston, MA.

Eduard Hovy and Elisabeth Maier. 1995. Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper.

Noriko Kando. 1999. Text structure analysis as a tool to make retrieved documents usable. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, pages 126–135, Taipei, Taiwan.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet - representation, visualization, application. In *Proceedings of LREC*, volume V, pages 1485–1491, Las Palmas.

Hagen Langer, Harald Lüngen, and Petra Saskia Bayerl. 2004. Towards automatic annotation of text type structure: Experiments using an XML-annotated corpus and automatic text classification methods. In *Proceedings of the workshop on XML-based richly annotated corpora (XBRAC) at the LREC 2004*, pages 8–14, Lissabon.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization.* MIT Press, Cambridge, MA.

Livia Polanyi, Martin van den Berg, and David Ahn. 2003. Discourse structure and sentential information structure. *Journal of Logic, Language and Information*, 12:337–350.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004a. A rule based approach to discourse parsing. In *Proceedings of the 5th Workshop in Discourse and Dialogue*, pages 108–117, Cambridge, MA. 2004.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004b. Sentential structure and discourse parsing. In *Proceedings of the ACL 2004 Workshop on Discourse Annotation*, pages 49–56, Barcelona.

David Reitter and Manfred Stede. 2003. Step by step: Underspecified markup in incremental rhetorical analysis. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at the EACL*, Budapest.

John M. Swales. 1990. *Genre Analysis. English in academic and research settings.* Cambridge University Press, Cambridge, UK.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71, Washington D.C. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Summarizing scientfic articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text.* Ph.D. thesis, University of Edinburgh.

Teun A. van Dijk. 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Norman Walsh and Leonard Muellner. 1999. *DocBook: The Definitive Guide.* O'Reilly.

Andreas Witt, Harald Lüngen, Daniela Goecke, and Felix Sasaki. 2005. Unification of XML documents with concurrent markup. *Literary and Linguistic Computing*, 20(1):103–116.