ARBEITEN ZUR MEHRSPRACHIGKEIT
WORKING PAPERS IN MULTILINGUALISM

96 • 2011

Folge B • Series B

Hanna Hedeland, Thomas Schmidt, Kai Wörner (eds.)

Multilingual Resources and Multilingual Applications

Proceedings of the Conference of the
German Society for Computational Linguistics and
Language Technology (GSCL) 2011

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Sonderforschungsbereich
Mehrsprachigkeit

Annotation of Explicit and Implicit Discourse Relations in the TüBa-D/Z Treebank

Anna Gastel, Sabrina Schulze, Yannick Versley, Erhard Hinrichs
SFB 833, Universität Tübingen

E-mail: (yannick.versley|erhard.hinrichs|sabrina.schulze)@uni-tuebingen.de, anna.gastel@student.uni-tuebingen.de,

Abstract

We report on an effort to add annotation for discourse relations, discourse structure, and topic segmentation to a subset of the texts of the Tübingen Treebank of Written German (TüBa-D/Z), which will allow the study of discourse relations and discourse structure in the context of the other information currently present in the corpus (including syntax, referential annotation, and named entities). This paper motivates the design decisions taken in the context of existing annotation schemes for RST, SDRT or the Penn Discourse Treebank, provides an overview over the annotation scheme and presents the result of an agreement study. In the agreement study, we use the notion of *inter-adjudicator agreement* to show that the task of discourse annotation, while challenging in principle, can be successfully solved when using appropriate heuristics.

Keywords: discourse, annotation, text segmentation, agreement

1. Introduction

Discourse information has been proven useful for a number of tasks, including summarization (Schilder, 2002) and information extraction (Somasundaran et al., 2009). While coreference corpora exist for many languages, and in large and very large sizes (frequently over one million words), the annotation of discourse structure and discourse relations has only recently gained the interest of the community at large.

Many of the existing corpora containing discourse structure and/or discourse relations are tightly bound to existing discourse theories such as Rhetorical Structure Theory (RST, Mann & Thompson, 1988) or Segmented Discourse Representation Theory (Asher, 1993), or subscribe to a fundament of coherence relations while avoiding assumptions about discourse structure (Hobbs, 1985; Wolf & Gibson, 2005).

While annotation guidelines for corpora such as the RST Discourse Treebank (Carlson et al., 2003; see Stede 2004, and van der Vlieth et al., 2011 for German and Dutch corpora, respectively, following these guidelines), an SDRT corpus (Hunter et al., 2007), or the Penn Discourse Treebank (PDTB, Prasad et al., 2007; see Al-Saif & Markert, 2010 for an effort towards an Arabic counterpart) generally agree on the idea of discourse relations between discourse segments, they do differ in other important aspects: RST (in particular, Carlson & Marcu, 2001) and the SDRT guidelines of (Reese et al., 2007) start from **elementary discourse units** (EDUs) that form the lowest level of a hierarchical structure; the PDTB's guidelines avoid the notion of discourse units, elementary or not, by asking annotators to mark **connective arguments** which may, but do not have to, coincide with syntactic or larger units, and do not need to form a hierarchy.

In terms of the relation inventory, the most important desideratum consists in reconciling descriptive adequacy for the linguistic phenomena involved with an inventory size that can still be annotated reliably. This problem is solved in different ways: The RST guidelines contain a coarse level of 16 relation classes, which are further specified into 78 relations which are organized by **nuclearity** (where mononuclear relations put greater weight on one of the units, the nucleus, whereas

CONTIGENCY [28.8%]
- **Causal** [20.5%]
  - (c)Result-Cause (5.9%)
  - (c)Result-Enable (4.7%)
  - (c)Result-Epistemic (0.4%)
  - (c)Result-Speechact (0.4%)
  - (s)Explanation-Cause (6.6%)
  - (s)Explanation-Enable (1.2%)
  - (s)Explantion-Epistemic (1.1%)
  - (s)Explanation-Speechact (0.6%)
- **Conditional** [3.0%]
  - (c)Consequence (2.1%)
  - (c)Alternation (0.5%)
  - (c)Condition (0.5%)
- **Denial** [5.6%]
  - (c)ConcessionC (4.0%)
  - (s)Concession (2.0%)
  - (s)Anti-Explanation (0.5%)

EXPANSION [43.6%]
- **Elaboration** [23.6%]
  - (s)Restatement (10.9%)
  - (s)Instance (3.4%)
  - (s)InstanceV (1.0%)
  - (s)Background (9.1%)
- **Interpretation** [4.2%]
  - (s)Summary (1.0%)
  - (s)Commentary (3.3%)
- **Continuation** [6.8%]
  - (c)Continuation (6.4%)

TEMPORAL [14.35%]
- (c)Narration (9.3%)
- (s)Precondition (2.4%)

COMPARISON [11,.%]
- (c)Parallel (3.3%)
- (c)ParallelV (1.1%)
- (c)Contrast (7.0%)

REPORTING [9.5%]
- (s)Attribution (4.2%)
- (s)Source (6.0%)

Table 1: Taxonomy of discourse relations with corpus frequencies

multinuclear relations connect units that are equally important); Reese et al's guidelines for SDRT annotation do not posit any larger categories among their 14 relations, but organize them by a distinction between **coordinating** and **subordinating** relations (cf. Asher & Vieu, 2005; this distinction vaguely corresponds to RST's notion of nuclearity), as well as by **veridicality** (where a relation is veridical if the larger unit containing it cannot be asserted without also asserting the truth of the relation arguments). The PDTB, in contrast, contains 30 relations which are organized into a taxonomy with 16 relations at the middle level and 4 relatively coarse top-level classes (Temporal, Contingency, Comparison, Expansion).

For someone aiming to annotate a corpus with discourse structure, the choice is not easy: The Penn Discourse Treebank carefully avoids any strong commitments to the ideas it uses as a backdrop (such as Webber 2004; Knott et al., 2001), treating the annotation more like a collection of examples that can be mined to verify aspects of the theory; Al-Saif and Markert (2010), for their work on PDTB-style annotation of Arabic discourse, found it necessary to drastically simplify the annotation scheme (from 30 to 12 relations) in order to yield a feasible scheme for their annotation of explicit discourse connectives.

Rhetorical Structure Theory, the most mature of the models for an annotation scheme, has also drawn a commensurate amount of (oftentimes valid) criticism: The most important one is that RST defines its relations in terms of speaker intentions, which yields good descriptive adequacy (given an appropriate inventory of relations), but fares less well for cognitive plausibility (cf. the overview of critiques in Taboada & Mann, 2006), with Sanders and Spooren (1999) claiming that RST lacks a separation between **intentions**, which are defined in terms of speaker and hearer, and their goals (as is customary in RST), and **coherence relations**, which connect two propositions. In a similar vein, Stede (2008) puts forward the claim that RST's notion of nuclearity encompasses criteria on different linguistic levels that are not always in agreement with each other.

Despite SDRT's focus on coherence relations and its strong theoretical commitment on coherence relations and their role in structuring the text, attempts to realize these principles in a general scheme for the discourse annotation of text have been few and far in-between, with the unpublished corpus of Hunter et al (2007) being the most notable example.

Hierarchical structuring of discourse is a well-established concept, not only because it reflects the principles that have been successful in structural accounts of syntax (see Polanyi & Scha, 1983; Grosz & Sidner, 1986, or Webber, 1991, *inter alia*), but also because it allows us to formulate well-formedness (coherence) constraints, as well as accessibility (Webber, 1991) in terms of local configurations.

While such a tree structure is classically motivated through intentional notions (the *discourse segment purposes* of Grosz & Sidner, 1986), the notion of *question under discussion* has been used in information structure to explain intonational focus in terms of (a hierarchy of) question under discussion (van Kuppevelt, 1995; Roberts, 1996; Büring 2003; also Polanyi et al., 2003 for a related proposal). It also allows to couch well-formedness in terms of valid sub-questions (for subordination) or being (non-exhaustive) answers to a common question (for coordination; cf. Txurruka, 2003).

Hence, we have, in addition to object-level relations (part-of, causality), an additional level of relations such as *Contrast* which are explainable in terms of information-structural notions, and which yet fulfill the intuition (made explicit by Roberts, 1996) that at any given point in discourse, interlocutors have a common notion of the discourse structure. This level is distinct from the upper-level structure that is the result of conscious structuring of the writer (possibly following genre-specific rules). As an example, some of the very general RST relations such as *Motivation* or *Preparation* are only explainable in terms of writer intentions and conscious text structuring, which may or may not be transparent to the average recipient.

Our own annotation scheme reflects van Kuppevelt's and Roberts' intuitions about a shared structure in discourse: We found it important to keep a backbone of explicit hierarchical structure, as in RST's annotation scheme, but also to avoid vague relations between large text segments, which are often genre-specific or the (sometimes idiosyncratic) result of intentional text structuring by the author. The PDTB successfully uses the metaphor of **implicit connectives** to limit discourse relations to connective-argument-sized pieces; in our case, we reconcile an explicit notion of (shallow) hierarchy with a focus on coherence relations by dividing the text into topically coherent stretches (as discussed, e.g., by Hearst, 1997), which we call **topic segments**, and annotate hierarchical discourse structure (using SDRT's notion of co- and subordinating discourse relations) inside these topic segments.

In the following text, section 2 gives more details on the corpus and on the annotation scheme, whereas section 3 presents an experiment to establish the reliability of our scheme using an inter-annotator agreement study. Section 4 presents and summarizes our findings.

2. Corpus and Annotation Scheme

As a textual basis for the corpus, we selected newspaper articles from the syntactically and referentially annotated TüBa-D/Z corpus (Telljohann et al., 2009), with the current version totalling 919 sentences in 31 articles, or about 29.6 sentences/article (against 20.6 sentences/article on average in the complete TüBa-D/Z, which also includes very brief newswire-style reports), and altogether 1159 discourse relations and 103 topic segments (or about 9 sentences per topic segment).

The relation inventory, and the distribution of different relation types, is presented in Table 1. From the starting point of the coordinating and subordinating discourse relations in Reese et al., we found it necessary to introduce finer distinctions in some places to ensure either consistency with a related effort on annotating explicit connectives (adding new relations such as *Result-enable* which corresponds to the *Weak-Result* relation proposed by Bras et al., 2006, for SDRT), but also the distinction between **Contrast** and **Concession** which is found in both the Penn Discourse Treebank and the RST annotation guidelines, but not Reese et al.'s proposal.

The resulting 28 relations can be grouped into 8 medium-level and 5 upper-level relation types by considering properties such as **basic operation** (causal vs. additive vs. temporal, with referential as a new group to account for elaborative relations) and **symmetry** as proposed by Sanders et al (1992); the resulting higher-level types of discourse relations have much in common with the top-level taxonomic categories of the Penn Discourse Treebank with a small number of exceptions (the PDTB subsumes the non-symmetrical *Concession* relation under the label *Comparison* whereas we follow Sanders et al. in assuming a **causal** source of coherence for *Concession* and an **additive** source of coherence for the symmetrical *Contrast* relation; Our **Reporting** group includes the *Attribution* and *Source* relations that Hunter et al. use in accounting for reported facts, whereas the Penn Discourse Treebank, unlike RST and SDRT, treats attribution as an issue that is orthogonal to discourse structure).

The hierarchical organization of relations according to basic operation does not differentiate between additional properties such as coordination/subordination or veridicality. Examples (1) and (2) serve to illustrate this distinction:[1]

(1) a) Private Unternehmen dürfen die Telefonbücher der Telekom-Tochter DeTeMedien nicht ohne deren Erlaubnis zur Herstellung einer Telefonauskunfts-CDs verwenden.
b) Die beklagten Unternehmen müssen den Vertrieb der Info-CDs sofort einstellen.

Result-Cause(1a,1b)

(2) a) Taxifahrer sind als Kolumnenthema eigentlich tabu,
b) weil sie als "weiche Angriffsziele" gelten.

Explanation-Cause(2a,2b)

When the situation specified in Arg1(1a) is interpreted as the cause of the situation specified in Arg2 (1b), the relation between those two arguments is labeled Result-Cause. Both arguments are necessary for coherence, so they are coordinated. The second example is labeled Explanation-Cause, because the situation specified in Arg1(2a) is interpreted as the result of the situation specified in Arg2 (2b). The situation in (2a) contains the main information while the situation in (2b) contributes background information. With subordinating relations, Arg2 ('further information') is always subordinated to Arg1 ('main information'), independently of surface order, as you can see in the following two examples:

(3) a) Zwei Ex-Mafiosi behaupten zudem,
b) von dem Mordauftrag Andreottis gewußt zu haben.

Attribution(3a,3b)

(4) a) Nach Angaben von Polizeipräsident Hagen Saberschinsky
b) haben Polizeibeamte einen ihrer Kollegen angezeigt.

Source(4b,4a)

In example (3) the main information is situated in Arg1: It is relevant for the coherence of the text to know that two mobsters testified knowing about the murder contract of Andreotti, which makes them important witnesses in the murder charges against Andreotti. Therefore Arg2 is subordinated to Arg1. In example (4) the main information, namely that police officers press charges against one of their colleagues, is given by (4b). Therefore, 4b is the Arg1 of a *Source* relation, as it is more important to know about the complaint itself than to know where the information came from, and 4a is subordinated under 4b (cf. Hunter et al., 2007).

Table 1 contains all discourse relations. Numbers in square brackets represent the distribution of the overall class. Numbers in parentheses represent the distribution of the single relation.

In the table, coordinating relations are marked with a small 'c' in front of the relation and subordinating relations are marked with a small 's'.

3. An experiment on inter-annotator and inter-adjudicator agreement

For any annotation scheme that ventures into the domain of semantic and/or pragmatic distinctions, reliability is an issue that needs to be addressed explicitly in order to maintain the predictability of the annotated data (or, equivalently, the predictive power of conclusions from that data).

Regarding the agreement on discourse relations, Marcu et al. (1999) determined κ values between κ=0.54 (Brown corpus) and κ=0.62 (MUC) for fine-grained RST relations and between κ =0.59 (Brown) and κ =0.66 (MUC) for coarser-grained relations. In their reliability study with the Penn Discourse Treebank, Prasad et al. (2008) determined agreement values between 80% (finest level) and 94% (coarsest level with 4 relation types), but did not report any chance-corrected values. Al-Saif and Markert (2010) report values of κ=0.57 for their PDTB-inspired connective scheme, saying that most disagreements are due to highly ambiguous connectives such as *w/and*, which can receive one of several relations. In a study on their Dutch RST corpus, van der Vlieth et al. (2011) found an inter-annotator agreement of κ=0.57. To the best of our knowledge, no agreement figures have been published on the RST-based Potsdam Commentary Corpus (Stede, 2004) or any other German corpus with discourse relation annotation.

In the regular annotation process of our corpus, two annotators create EDU segmentation, topic segments,

[1] TüBa-D/Z sentences 2563/2564, 7482/7483

and discourse relations independently from each other; in a second step, the results from both annotators are compared and a coherent gold-standard annotation is created after discussing the goodness-of-fit of respective partial analyses to the text and the applicability of linguistic tests. In order to account for the complete annotation process including the revision step, we follow Burchardt et al. (2006) and separately report *inter-annotator* agreement, which is determined after the initial annotation, and *inter-adjudicator* agreement, which is determined after an additional adjudication step. The adjudication step is carried out by two adjudicators based on the original set of annotations, but is performed by each adjudicator independently from the other.

In the case where multiple relations were annotated between the same EDU ranges (for example, a temporal *Narration* relation in addition to a *Result-Cause* relation from the Contingency group), we counted the annotations as matching whenever the complete set of relations (i.e. {*Narration, Result-Cause*} in the example) is the same across annotators.

In a sample of three documents that we used for our agreement study, we found that annotators agreed on 49 relations spans, with the comparison yielding an agreement value of κ=0.55 for individual relations, and κ=0.65 for the middle level of the taxonomy (eight relation types).

For the inter-adjudicator task, we found an agreement on 82 relation spans, among which relation agreement was at κ=0.83 for individual relations, and κ=0.85 for the middle level of the taxonomy, or a reduction of disagreements of about 57%.

4. Discussion and Conclusion

In this article, we have presented the annotation scheme we use to annotate discourse relations of complete texts in a subset of the TüBa-D/Z corpus, and reported the results of an agreement study using these guidelines and relation inventory. While the raw inter-annotator agreement is on a similar level as other annotation efforts with a similar scope, we found that a subsequent adjudication step introduces a rather substantial reduction in disagreements (between adjudicated versions that were obtained independently of each other), which suggests that a large part of the (raw) disagreement is due to the sheer complexity of the task and should not be taken as indicating the infeasibility of discourse structure (and discourse relation) annotation in general.

The public availability of a corpus with discourse relation annotation in combination with the syntactic and referential annotation from the main TüBa-D/Z corpus will also allow it to provide an empirical evaluation of theories concerning the interface between syntax and discourse, such as D-LTAG (Webber, 2004) or D-STAG (Danlos, 2009) as well as those that predict interactions between referential and discourse structure (Grosz & Sidner 1986; Cristea et al., 1998; Webber, 1991; Chiarcos & Krasavina, 2005, inter alia).

5. References

Al-Saif, A., Markert, K. (2010): Annotating discourse connectives for Arabic. In Proc. LREC 2010.

Asher (1993): Reference to Abstract Objects in Discourse. Kluwer, Dordrecht.

Asher, N., Lascarides, A. (2003): Logics of Conversation. Cambridge University Press, Cambridge.

Asher, N., Vieu, L. (2005): Subordinating and coordinating discourse relations. Lingua 115, 591-610.

Bras, M., Le Draoulec, A., Asher, N. (2006): Evidence for a Scalar Analysis of Result in SDRT from a Study of the French Temporal Connective 'alors'. In: SPRIK Conference "Explicit and Implicit Information in Text - Information Structure across Languages".

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., Pinkal, M. (2006): The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In Proceedings of LREC 2006.

Büring, D. (2003): On D-Trees, Beans, and B-Accents. Linguistics and Philosophy 26(5), pp. 511-545.

Carlson, L., Marcu, D. (2001): Discourse Tagging Manual. ISI Tech Report ISI-TR-545.

Carlson, L., Marcu, D., Okurowski, M. E. (2003): Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In: Current Directions in Discourse and Dialogue, Kluwer.

Chiarcos, C., Krasavina, O. (2005): Rhetorical Distance

Revisited: A Parametrized Approach. In Workshop on Constraints in Discourse (CID 2005).

Cristea, D., Ide, N., Romary, L. (1998): Veins Theory: A Model of Global Discourse Cohesion and Coherence. In Proc. CoLing 1998.

Danlos L. (2009): D-STAG : Un formalisme d'analyse automatique de discours basé sur les TAG synchrones. Revue TAL 50 (1), pp. 111-143.

Grosz, B., Sidner, C. (1986): Attention, Intentions, and the structure of discourse. Computational Linguistics 12(3), pp. 175-204.

Hearst, M. (1997): TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, Computational Linguistics, 23 (1), pp. 33-64.

Hobbs, J. (1985): On the Coherence and Structure of Discourse, Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hunter, J., Baldridge, J., N. Asher (2007): Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts. Zeitschrift für Sprachwissenschaft 26, pp. 213-239.

Knott, A., Oberlander, J., O'Donnell, M., Mellish, C. (2001): Beyond Elaboration: The interaction of relations and focus in coherent text. In: Sanders, Schilperoord, Spooren (eds.), Text representation: linguistic and psycholinguistic aspects. John Benjamins.

Mann, W. C., Thompson, S. A. (1998): Rhetorical Structure Theory: Toward a functional theory of text organization. Text 8, pp. 243-281.

Marcu, D., Amorrortu, E., Romera, M. (1999): Experiments in Constructing a Corpus of Discourse Trees. ACL Workshop on Standards and Tools for Discourse Tagging.

Polanyi, L., Scha. R. (1983): On the Recursive Structure of Discourse. In K. Ehlich & H. Van Riemsdijk (Eds.), Connectedness in sentence, discourse and text, pp. 141–178. Tilburg: Tilburg University

Prasad, R., Miltsakaki, M., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., Webber, B. (2007): The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report, University of Pennsylvania.

Reese, B., Denis, P., Asher, N., Baldridge, J., Hunter, J. (2007): Reference Manual for the Analysis and Annotation of Rhetorical Structure. Technical Report, University of Texas at Austin.

Roberts, C. (1996): Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In Yoon, Kathol (eds.), OSU Workin Papers in Linguistics 49: Papers in Semantics, pp. 91-136.

Sanders, T. J. M., Spooren, W. P. M., Noordman, L. G. M. (1992): Toward a Taxonomy of Coherence Relations. Discourse Processes 15, pp. 1-35.

Sanders, T. J. M., Spooren, W. P. M. (1999): Communicative intentions and coherence relations. In Bublitz, Lenk, Ventola (eds.) Coherence in Text and Discourse, pp. 235-250. John Benjamins, Amsterdam.

Schilder, F. (2002): Robust discourse parsing via discourse markers, topicality and position. Natural Language Engineering 8(2), pp. 235-255.

Somasundaran, S., Namata, G., Wiebe, J., Getoor, L. (2009): Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification. In Proc. EMNLP 2009.

Stede, M. (2004): The Potsdam Commentary Corpus. In Proc. ACL Workshop on Discourse Annotation.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., Beck, K. (2009): Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical Report, Seminar für Sprachwissenschaft, Universität Tübingen.

Txurruka, I. G. (2003): The Natural Language Conjunction And. Linguistics and Philosophy 26(3), pp. 255-285.

van der Vlieth, N., Berzlanovich, I., Bouma G., Egg, M., Redeker, G. (2011): Building a Discourse-Annotated Dutch Text Corpus. In Proceedings of the DGfS Workshop "Beyond Semantics", Bochumer Linguistische Arbeitsberichte 3.

van Kuppevelt, J. (1995): Discourse Structure, Topicality and Questioning. Linguistics 31, pp. 109-147.

Webber, B. (1991): Structure and Ostension in the Interpretation of Discourse Deixis. Natural Language and Cognitive Processes 6(2), pp. 107-135.

Webber, B. (2004): DLTAG: Extending Lexicalized TAG to Discourse. Cognitive Science 28, pp. 751-779.