# ANNODIS and related projects: case studies on the annotation of discourse structure

Nicholas Asher, Philippe Muller, Myriam Bras, Lydia-Mai Ho-Dac,
Farah Benamara, Stergos Afantenos and Laure Vieu

March 16, 2015

**Abstract**

In this paper we report on the efforts of three projects to annotate texts and dialogues with discourse structure. We provide a theoretical discussion of various alternatives and then present our approach to discourse structure annotation, along with some applications of the resources that we have developed.

## 1 Introduction

It is a commonplace that texts and conversations are not just bags of sentences, just as sentences are not just bags of words. Like sentences, discourses have structure in which discourse constituents may play one or more discursive roles. In the words of Webber *et al.* (2012): "Discourse structures are the *patterns* that one sees in multi-sentence (multi-clausal) texts. Recognizing these pattern(s) in terms of the elements that compose them is essential to correctly deriving and interpreting information in the text." Most researchers working on discourse would also maintain that a well-formed discourse structure is essential for discourse coherence. Previous work over the last 20 years has demonstrated that this discourse structure has important effects on the content that competent interpreters glean from texts in a variety of areas—anaphora, ellipsis, temporal structure and lexical disambiguation (Hobbs, 1979; Lascarides & Asher, 1993; Hitzeman *et al.* , 1995; Asher *et al.* , 2001; Asher, 2011). Discourse structure is thus an important component for calculating the overall meaning of a text or conversation. Given this, the extraction of discourse structure from texts has many applications, among which are text summarization, information retrieval, question answering, sentiment or opinion analysis. In this chapter, we provide a discussion of our efforts to annotate discourse structure in text and some of the applications to which these annotations have been put.

## 2 Theoretical preliminaries

This chapter provides a case study of annotation for discourse structure. As we have said, it is widely agreed that discourse structure affects the interpretation or meaning of a text. But beyond that, there are some theoretical choices. Most linguists would accept some version of Montague's homomorphism from syntactic structures to semantics. Moving to the textual level, the question is where do we introduce discourse

structure? Is it an extension of the syntactic component of language; i.e., is it an extension of a syntactic parse or parses of a text's constituent sentences? Or is it rather an extension of the semantic component which takes syntactic parses and converts them into semantically transparent representations for which can be defined a notion of logical consequence, and hence a mechanism for predicting semantic entailments? Most work on discourse structure would take the latter position, though to some extent it is a matter of taste. In so doing, they take semantic representations, propositions (Forbes *et al.* , 2003), occurrences of propositions (Asher, 1993) or some other semantic entity as the *relata* of *discourse relations*, which themselves are semantically defined in terms of what content they add to the text. A discourse structure then is a semantic object, a graph involving some sort of semantic entities as vertices and a relational structure over those entities. Discourse theorists who do not develop a formal approach to discourse structure also mainly subscribe to this view (Halliday, 1977).

This choice has of course an effect on the design of the annotation and the annotation manual: discourse relations or structures are defined in semantic terms, and a wide choice of features, syntactic and presentational (e.g. having to do with a text's layout) but also semantic features like verb classes or lexical classes generally, *aktionsart*, the presence of anaphors, etc. can be exploited in determining the nature of the discourse structure.

The next choice point has to do with the nature of the discourse structure one wants to investigate. Does one want to investigate the discourse structure of the whole text—i.e., is the object of study a connected graph, in which every relevant semantic entity is linked to some other entity in the structure for a coherent text? Alternatively, one may study the occurrence of just selected kinds of structures in a text, ones for instance that are linked to certain features. One example of such a structure, discussed in section 3.4, is what we call an *enumerative structure*, and it has a special list of features and structure all its own. The second annotation campaign we discuss below features both annotations for the discourse structure of a whole text and annotations of one particular sort of structure across a wide range of texts.

To this question, we add another. Given that one wants to study such structures, how are they to be defined? Most theories on the market—Rhetorical Structure Theory (RST) (Mann & Thompson, 1987), the Linguistic Discourse Model (LDM) (Polanyi *et al.* , 2004), the GraphBank model (Wolf & Gibson, 2005), Discourse Lexicalized Tree Adjoining Grammar (DLTAG) (Forbes *et al.* , 2003), the Penn Discourse Treebank model (PDTB) (Prasad *et al.* , 2008), and Segmented Discourse Representation Theory (SDRT) (Asher, 1993) define hierarchical structures by constructing complex discourse units (CDUs) from elementary discourse units (EDUs), i.e., "bottom-up", in recursive fashion. This follows standard practice when defining logical languages and providing their semantics.

Alternatively, one might construct either a partial or full discourse structure in "top-down" fashion, which starts by finding the representation of a text's macro-organization. This "top-down approach" focuses on "multi-level" text spans and signals of global text organization (Enkvist, 1989; Chafe, 1994; Fries, 1995; Goutsos, 1996; Power *et al.* , 2003; Ho-Dac & Péry-Woodley, 2009).

The top-down and bottom-up approaches can give equivalent results (as is well-known for the construction of semantic representations like those in DRT (Kamp & Reyle, 1993)), but they typically emphasize different parts of discourse structure. The top-down perspective suggests that readers perceive (or believe in) the text's coherence before constructing their interpretation unit by unit and they detect large scale structures before detailing the lower level aspects of the complete discourse structure for a

text. Goutsos (1996) for instance takes the detection of continuities and discontinuities as fundamental. From the relational perspective, this means looking at chunks that are individuated by a lack of local attachments; i.e., the chunks are attached higher up to some other constituent or to each other but no links occur between elements of those chunks.

Although Goutsos considers only thematic (dis)continuity, we argue that the specific interpretation criteria which bind text units together into larger units may concern different levels of organization: thematic continuity but also space/time reference, the presence of a particular rhetorical or discourse structure in the sense of the bottom up approach, as well as the typographical presentation of the text itself. A shift between two segments may be a referential break, the end or opening of a discourse frame, or the end or beginning of a paragraph or a section. Detecting discontinuities or what is known as discourse pops from the bottom up perspective is often quite difficult, as we will detail below; so in principle, such top down criteria can be complementary to those given by a bottom up approach. The annotation campaign of ANNODIS featured both a bottom up and top down approach to discourse annotation.

## 2.1 Recursive and complete discourse structures for text and dialogue

Let us suppose that the object of study is a complete discourse structure for a text or dialogue, in which every constituent is linked to some other constituent.[1] Both top-down and bottom up strategies share certain tasks: the bottom up approach needs to decide *where to start*—i.e., what are the basic or elementary discourse units, while the top down approach needs to decide *where to stop*—i.e. at what point discourse structure ends and clause level semantics begins; the bottom up approach needs to decide how to combine elementary units together to build larger ones, while the top-down approach needs to decide how to break larger structures down into smaller ones; finally both approaches need to decide how to link discourse constituents—i.e. what are the relations that bind distinct discourse units into a coherent whole. Thus, to get a complete structure for a text three decisions need to be made:

- what are the elementary discourse units or constituents (EDUs)
- how do elementary units combine to form larger units and attach to other units?
- how are the links between discourse units labelled with discourse relations?

We believe these questions are best answered in the context of an awareness of theoretical frameworks for the analysis of discourse and discourse interpretation. These frameworks have developed answers to these questions and often offer a coherent picture of what discourse structure is and what it does to interpretation. This theoretical work can save designers of discourse annotation schemes from making choices that we know to be wrong or very unpromising. That said, annotation scheme designers have to weigh what this theoretical work says with respect to what sort of annotation they want to do: some choices proposed by some theories may be suitable for some annotation tasks and not for others. We try to highlight some examples of data confronting theory below.

**Elementary Discourse Units:** theories and annotation schemes have contributed different answers concerning the nature of EDUs. Many theories (RST, DLTAG) take

[1] Not all annotation campaigns of course have this as a goal, the PDTB being one prominent example.

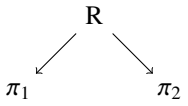
full sentences or at least tensed clauses as the mark of an EDU. SDRT, as developed in (Asher & Lascarides, 2003) was largely mute on the subject of EDU segmentation, but in general also followed this policy. A detailed examination of the semantic behavior of appositives, non restrictive relative clauses and other parenthetical material in our corpora, however, revealed that such syntactic structures also contributed EDUs. Such constructions provide semantic contents that do not fall within the scope of discourse relations or operators between the constituents in which they occur. For example, in (2.1), we see that appositions do not or at least need not fall within the scope of the conditional or the attribution relation on a defensible interpretation of the text. This semantic behavior indicates that the contents contributed by such constructions are not to be treated as part of the tensed clauses in which they occur.

**Example 2.1.**

If the former President of the United States, *who has been all but absent from political discussions since the 2008 election*, were to weigh in on the costs of the economic shutdown, the radical Republicans might be persuaded to vote to lift the debt ceiling.

A spokesman said that Steven Jobs, *the CEO of Apple*, would address stockholders at the upcoming shareholder's meeting.

**Attachment decisions:** There is a divide between those discourse frameworks that take discourse structure to be trees (DLTAG, LDM, RST) and those that take discourse structures to be some sort of non-tree-like graph (SDRT, Graphbank). There are at least two parameters that influence this decision. The first is: should the discourse annotations or the discourse structures that result from the annotation process make explicit the semantic scope for the discourse relations—e.g., should an RST-like structure, in which leaves are EDUs and all non terminal nodes are labelled with discourse relations, like




have the natural interpretation that the relation $R$ has as its left argument the constituent $\pi_1$ and as its right argument the constituent $\pi_2$? If the structures are trees and the natural interpretation is the one adopted, then one has trouble making sense of long distance attachments. While this immediate interpretation is standard in SDRT, it is not in RST. Consider the examples in (2.2, taken from the RST Tree Bank and the main corpus described here, and from the ANNODIS corpus (Afantenos *et al.* , 2012), discussed in (Venant *et al.* , 2013):

**Example 2.2.**

a) [In 1988, Kidder eked out a $ 46 million profit,]$_{31}$ [mainly because of severe cost cutting.]$_{32}$ [Its 1,400-member brokerage operation reported an estimated $ 5 million loss last year,]$_{33}$ [although Kidder expects to turn a profit this year]$_{34}$ (RST Treebank, wsj_0604).

b) [Suzanne Sequin passed away Saturday at the communal hospital of Bar-le-Duc,]$_3$ [where she had been admitted a month ago.]$_4$ [She would be 79 years old today.]$_5$ [...] [Her funeral will be held today at 10h30 at the church of Saint-Etienne of Bar-le-Duc.]$_6$ (ANNODIS corpus, ER045, English translation).

These examples involve what are called *long distance attachments*. Example 2.2-a involves a relation of contrast, or comparison between 31 and 33, but which does not involve the contribution of 32 (the costs cutting of 1988). Example 2.2-b displays something comparable. A causal relation like result, or at least a temporal narration holds between 3 and 6, but it should not scope over 4 and 5 if one does not wish to make Sequin's admission to the hospital a month ago and her turning 79 a consequence of her death last Saturday.

It is impossible however, to account for such long distance attachment using the immediate interpretation of RST trees. 2.2-a, for instance, also involves an explanation relation between 31 and 32, which should include none of 33 or 34 in its scope. Since 31 is in the scope of both the explanation and the contrast relation, an RST tree involving the two relations has to make one of the two relations dominate the other in the tree representation.

To handle such difficulties, researchers have explored two options. The first is to develop a non immediate interpretation of an RST structure, which typically involves another layer of annotation in which some nodes are labelled *nucleus* and others labelled *subordinate*. This additional layer of annotations is then used to compute the actual semantic scopes of discourse relations (see Marcu, 1996; Danlos, 2008; Egg & Redeker, 2010). The other option is to adjust the conception of the discourse structure so that the immediate interpretation is retained, as is done in SDRT. We have followed the second option in our annotation development.

**Types of Discourse Relations:** While theories and annotation schemes differ to some extent on what types of discourse relations there are, a consensus has emerged on a general typology for written texts. Most annotation models include relations that allow for various kinds of expansion or elaboration of a given discourse unit, explanatory links (why an event described in one discourse unit occurred), narrative and forward causal sequences, and structural relations like Parallel and Contrast. However, the characterization of a unique set of relations both suitable to accurately describe all attachments in a corpus, and of a size and granularity appropriate for this part of the annotation task remains a controversial and difficult task. Part of the problem is that the characterization of such relations is often vague and varies in much of the literature. SDRT insists on a semantic characterization of relations, which provides a method of verifying whether two relations are the same, one entails the other, are independent or are incompatible. We have used this approach in our annotation manual (see below) to describe a relation independently from its possible discourse markers, too often ambiguous, and to focus on what distinguishes relations that are often confused.

When we move from texts to dialogues, though the discourse structure of dialogues has received less attention with respect to formal modeling (*pace* Grosz & Sidner (1986)), we cannot just use a set of relations that are adequate for characterizing attachments in texts. In dialogue, questions and special relations involving them are pervasive Carletta *et al.* (1996). In addition dialogue features relations that encode disagreements and agreements between speakers. We have found that the discourse relations used to label attachments for dialogue will be a superset of those in monologue.

# 3 From model and raw data to annotation

## 3.1 DISCOR: a first experiment on discourse structure

A first effort on the part of some of the authors of this paper to build a annotated corpus with rhetorical relations was an NSF funded project, DISCOR (Baldridge *et al.* , 2007), carried out at the University of Texas at Austin. The project annotated 60 English texts from the MUC 6 and MUC 7 data sets, and so the texts were largely news stories. We used SDRT as the basis for our annotation model, and only experts in the theory did the annotation. We were quite naive and did the annotation by hand, beginning with EDU segmentation and then building the discourse structure from them. By and large, we found this to be a difficult and error-prone process and we came quickly to realize that more than one discourse annotation might be plausible given the cues present in the text. In particular standard measures of measuring agreement between annotators might have to be re-evaluated in this more semantic setting. Discourse pops and long distance attachments often gave rise to disagreements. On the other hand, we saw that the theory could be applied to open domain texts without too much difficulty, and annotator agreement for simple or short texts was often quite high. This gave us hope that perhaps we could build a bigger annotation campaign with less expert annotators.

## 3.2 ANNODIS: A second annotation campaign

Our next annotation effort, in which all of the authors of this chapter participated, attempted to come to grips with the annotation process in a more disciplined way. We investigated both the top-down and the bottom-up approaches to annotation on a corpus of French texts. As a result, we developed two annotation models with some common characteristics in order to bring the two closer and permit annotation comparison. The project, in particular the team from Caen involving Patrice Enjalbert, Antoine Widlöcher and Yann Mathet, also developed an annotation tool, Glozz (Mathet & Widlöcher, 2009)[2], specially designed for this purpose. Glozz is a generic annotation tool that allows one to annotate units, relations and schemes plus display texts with their visual typography— paragraph breaks, headings, bullets/numbered lists, etc. It also provides for the possibility for highlighting premarked features in order to assist annotation procedures.

Another common requirement was to take into account a diversified corpus, with a variety of genre, length and type of discursive organization. Nevertheless, while an annotation of rhetorical relations, that must be exhaustive, was inconceivable on long texts (e.g. academic papers), multi-level structures annotation needs long structured texts with multi-level headed section. As a result, the ANNODIS corpus was divided in two parts, corresponding for the bottom-up approach of short texts (a few hundred words each) and excerpts from longer documents and for the top-down approach, of longer (several thousands words each), complete and more complex documents. A small part of the corpus was annotated with both rhetorical relations and multi-level structures. Table 1 gives an overview of the ANNODIS corpus and the amount of annotated data. Five subcorpora are distinguished, issued from four different sources: NEWS (short news articles from the daily *Est Républicain*, publicly available), WIK1 (short excerpts of encyclopedia articles from the French Wikipedia), WIK2 (full encyclopedia articles from the French Wikipedia), LING (linguistics research papers from *CMLF: Colloque Mondial de Linguistique Française*) and GEOP international relation

[2] `http://glozz.free.fr/`

reports (from *IFRI: Institut Français des Relations Internationales*). Table 1 also distinguishes different types of annotated data with a breakdown by approach: on the one hand there are segmented elementary discourse units (EDU), rhetorical relations between units (Rh.Rel.) and complex discourse units (CDU) created; on the other hand, two multi-level structures: enumerative structures (ES) and topical chains (TC). These annotated data are described in the next subsections.

| corpus | Annotated objects | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | bottom-up approach | | | top-down approach | |
| | words | texts | EDU | Rh.Rel. | CDU | ES | TC |
| NEWS | 9,768 | 39 | 1159 | 1203 | 510 | | |
| WIK1 | 17,330 | 42 | 1949 | 2034 | 829 | | |
| WIK2 | 231,000 | 30 | 53 | 65 | 38 | 401 | 266 |
| LING | 169,000 | 25 | 12 | 14 | 9 | 297 | 88 |
| GEOP | 266,000 | 32 | 15 | 19 | 9 | 293 | 234 |
| ANNODIS | 687,000 | | 3188 | 3355 | 1395 | 991 | 588 |

Table 1: **Rhetorical relations and multi-level structures in the ANNODIS resource.** EDU = Elementary Discourse Units ; Rh.Rel. = Rhetorical Relations ; CDU = Complex Discourse Units ; ES = Enumerative Structures ; TC = Topical Chains.

Both approaches used the Glozz annotation platform for annotation: delimited units (elementary discourse units, coreferential expressions, enumerative structures components) are linked with specific (rhetorical) relations and grouped in schemas (complex discourse units, topical chains or enumerative structures). Secondly, the same process was followed: a first draft of the annotation manual was experimented by each other approach (top-down / bottom-up) and progressively modified. Both annotation manuals were then made into technical reports (Muller *et al.* , 2012b) and (Colléter *et al.* , 2012). The annotation procedure was more or less the same: on the basis of an annotation manual, three undergraduate students with no background in discourse theory or annotation practice annotated objects in texts by using the same tool (Glozz). For annotating multi-level structures, annotators started from a bird's eye view of texts and zoomed on specific zones. As for rhetorical relations, annotators started by segmenting texts into EDUs and, after mutual agreement, linked them with discourse relations and constructed CDUs in order to obtain a complete hierarchical representation of the text.

## 3.3 The bottom up approach in ANNODIS

Like the DISCOR project, the bottom-up approach in ANNODIS focused on providing a complete structure of a text, starting from the segmentation into EDUs (mostly clauses, appositions, some adverbials). Having learned from the DISCOR campaign, we spent a great deal of time developing an annotation manual for ANNODIS. Almost the first year of the project was devoted to annotation exercises between experts and a discussion of the results. Starting from the DISCOR/SDRT relation set, we decided to merge certain relations that proved difficult for experts to detect reliably (for example the distinction between two ways of annotating attributions in DISCOR) and introduced others, in particular a new sub-species of elaboration, entity-elaboration (Prévot *et al.* , 2009), to account for appositions as shown in the example above. We also used a "Frame" relation, which relates a framing adverbial and EDUs within its scope (Charolles, 1997): e.g. for *[During the 20th century,]*$_1$ *[EDU1]*$_2$. *[EDU2]*$_3$, we have

Frame(1,2) and Frame(1,3). The remaining relations chosen for linking discourse units were ones that are more or less common to all the theories of discourse, as mentioned above, or correspond to well-defined subgroups in fine-grained theories (Hovy, 1990). This intermediate level of granularity was chosen as a compromise between informativeness and reliability of the annotation process. It corresponds to the level chosen in the PDTB (see Part II, IV.b.i, this volume), and a coarse-grained RST. Our earlier work on these relations was helpful in detailing how these relations are linguistically marked in the annotation manual. The relations were each defined in semantic terms in the manual; for this we relied heavily on prior work mostly in the SDRT framework. The manual used the semantics to provide an intuitive idea for each relation, suitable for the level of the annotators. Occasional examples were provided. We gave a list of possible markers for each relation but we cautioned that the list was not exhaustive and that the markers were possibly ambiguous. Finally, we also made clear that a relation could occur in the absence of a marker or in spite of a marker that ordinarily signaled a different relation (for more details see section 4.1). The linguistic cues include not only so-called discourse markers but also tense and aspectual shifts, as well as specific syntactic structures. The relations used were the following: EXPLANATION, GOAL, RESULT, PARALLEL, CONTRAST, CONTINUATION, ALTERNATION, ATTRIBUTION, BACKGROUND, FLASHBACK, FRAME, TEMPORAL-LOCATION, ELABORATION, ENTITY-ELABORATION, COMMENT.

We also spent a long time developing guidelines for the segmentation of text into EDUs, which had not been done before to our knowledge, and which we incorporated into the annotation manual. The annotation manual provided annotators with an intuitive introduction to discourse segments, including the fact that we allowed discourse segments to be embedded in one another. Detailed instructions were then provided describing how to handle segmentation for most of the cases that could naturally arise, such as: simple phrases; conditional and correlative clauses; temporal, concessive or causal subordinate phrases; relative subordinate phrases; clefts, appositions, adverbials; coordinations, etc.

We then had a several month long trial period involving two graduate students in linguistics (who had little to no knowledge of theories of discourse structure), in which we iterated revisions on the annotation manual after examining the student annotations and discussing them. The two graduate-level students doubly annotated 50 documents. We built and regularly updated a wiki to keep track of our decisions concerning segmentation, discourse relations, and overall structures. This phase was extremely useful to us in detecting inconsistencies and incompletenesses in the manual. We also verified interannotator agreement between our subjects here and were confident enough with the results to begin our annotation campaign in earnest.

The bottom-up approach used both naive and expert annotators for the annotation campaign. The three undergraduate students doubly annotated 86 documents. They were trained for a week, with the help of the aforementioned manual and the graphical annotation tool Glozz. They segmented the texts into EDUs and adopted an agreed on segmentation, which Glozz then displayed to them for the next stage of the annotation process in which they introduced relations between EDUs. They were also given the possibility of creating larger scale structures, or complex discourse units (CDUs), if they wished to do so, using a schema template provided in Glozz. Over a period of one month of intensive annotation, the three students each annotated 2/3 of the corpus to produce a double annotation over 86 texts. Experts then adjudicated the annotations, often re-annotating close to from scratch, in particular when naive annotations were wrong or too distant.

The reason for the re-annotation had to do with a conscious choice concerning the

design of the annotation manual. We intentionally restricted the amount of information about discourse structure in the manual. It focused essentially on two aspects of the discourse annotation process: segmentation and typology of relations. Crucially, the manual did not provide any details concerning the structural postulates of the underlying theory. More specifically, we did not mention anything concerning distance of attachment, crossed dependencies and more theoretical postulates, such as constraints on attachment (the so-called "right frontier" of discourse structure), see section 4.1). We did this because we wanted to test the intuitions of the naive annotators relevant to these issues. We did mention, however, that whenever the annotators felt that strong coherence existed between a group of EDUs, they could lump them together in order to create a CDU which could then be linked with another EDU or CDU. We did not provide any further details on the nature of this coherence. An example of discourse, where CDUs are also included, is shown in example 3.1 translated from the ANNODIS resource.

It is not easy to define inter-annotator agreement on a relational task, as was done in ANNODIS, as opposed to annotation of isolated instances. We thus evaluated first the agreement on attachment decisions (which pairs of segments are related), and then the agreement on labels for segment pairs that were related by both annotators of the same text. We also considered as equally attached pairs of segments in any order, since a lot of errors were made on the order of arguments; we assume this was mostly because the annotation tool lacked ergonomic features needed for exhaustive text annotation—exhaustive annotation ended up cluttering the workspace making the end result very difficult to read. One of the three naive annotators was also very different from the other two, and we detail here only the best pair, pre-adjudication. These annotators agreed at 66% on attachments (taking the harmonic mean of both coverages, annotator 1 with respect to annotator 2 and vice versa). Kappa (Cohen, 1960) on the labels was 0.40, a moderate agreement according to the scale by (Landis & Koch, 1977). No transitivity of relations was assumed. It is noteworthy that some structures could be described differently from a "syntactic" annotation point of view, but corresponded to obviously equivalent structures from a semantic point of view; e.g., Elaboration (a,b) and Continuation (b,c) are semantically equivalent given our background assumptions to Elaboration(a,[b,c]), with [b,c] as a CDU). For lack of an explicit model of these equivalences, however, we could not account for these equivalences[3], and the raw agreement presented here is probably underestimated. Nonetheless, it prompted the expert annotation that yielded the final annotation[4].

Table 1 shows the number of EDUs, CDUs and rhetorical relations annotated in the corpus, with a breakdown by sub-corpus. Table 2 shows a breakdown of the relation types found in the corpus for the bottom-up approach. Information on the inter-annotator agreement is presented below.

## 3.4 Multi-level Structures annotation

As described in section 2 the concern of the top-down approach is with text organization strategies, viewed in a Systemic Functional framework (Halliday, 1985), and in particular with strategies regarding textual continuity and discontinuity (Goutsos, 1996). To translate this view into a realistic annotation program, an annotation model was devised focusing on the detection of two discourse structures highlighting the con-

[3]But see (Roze, 2013) for an investigation of some of these cases.

[4]see Part I, V.f, this volume, for a discussion on this point

**Example 3.1.**

[Milutinovic before the TPI.]_1[The former president of Serbia Milan Milutinovic, [accused along with the Yugoslav ex-head of State Slobodan Milosevic for war crimes in Kosovo,]_3 yesterday voluntarily turned himself over to the International Criminal Court for Ex-Yugoslavia in The Hague]_2 [Having arrived in the Netherlands in a plane of the Yugoslav government,]_4 [M.Milutinovic was emprisoned at the detention center of the Criminal Court at the beginning of the afternoon]_5

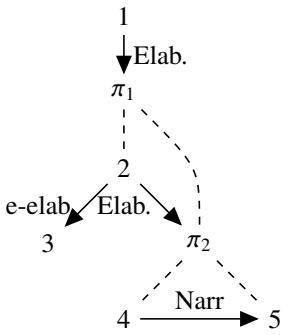



Figure 1: **An example of discourse graph**. The nodes correspond to discourse units; the EDUs are represented by their numbering; the CDUs start with $\pi$. Dotted edges represent inclusion to a CDU while edges with arrows represent rhetorical relations. Elab. = Elaboration, e-elab = Entity Elaboration, Narr. = Narration.

| | Nb | (%) | News % | Wik1 % | | Nb | (%) | News % | Wik1 % |
|---|---|---|---|---|---|---|---|---|---|
| alternation | 18 | 0.5 | 0.3 | 0.6 | explanation | 130 | 3.9 | 4.4 | 3.3 |
| attribution | 75 | 2.2 | 3.0 | 1.7 | flashback | 27 | 0.8 | 1.4 | 0.6 |
| background | 155 | 4.6 | 5.2 | 4.8 | frame | 211 | 6.3 | 6.2 | 5.7 |
| comment | 78 | 2.3 | 3.6 | 1.3 | goal | 95 | 2.8 | 3.1 | 2.4 |
| continuation | 681 | 20.3 | 20.1 | 21.1 | narration | 349 | 10.4 | 11.1 | 10.4 |
| contrast | 144 | 4.3 | 3.7 | 4.6 | parralel | 59 | 1.8 | 2.2 | 1.8 |
| E-elab | 527 | 15.7 | 14.1 | 16.4 | result | 163 | 4.9 | 4.7 | 5.4 |
| elaboration | 625 | 18.6 | 16.3 | 19.4 | temploc | 18 | 0.5 | 0.5 | 0.5 |
| totRel(nb) | 3355 | | | | | | | | |

Table 2: Discourse relations of the expert annotations

tinuity/discontinuity dichotomy: topical chains and enumerative structures.

Topical chains (TCs) are a specific type of cohesive chain (Halliday & Hasan, 1976): topically homogeneous segments, i.e. segments made up of sentences containing topical co-referential expressions. They may contain sentences which are not topically connected (e.g. comments, illustrations, etc.) if they occur between connected units.

Enumerative structures (ESs) are segments (in effect CDUs) consisting of three sub-segments: an optional **trigger** announcing the enumeration; several **items** composing the enumeration (at least two items); an optional **closure** which summarizes and/or closes the enumeration. Lexical expressions specifying the co-enumerability criterion are often present in the trigger and/or the closure. In the example 3.2, "*important groups*" is such an expression. Such an expression is boxed in the example 3.2. This example gives a text span translated from ANNODIS resource containing 1 ES detailing "*three important groups*" developed by *Saddam Hussein's regime* which constitutes the topic of 2 TCs. Topical expressions are italicized, ES cues are in bold and horizontal plain lines represents paragraph breaks.

Because enumerative structures typically come with a variety of clear cues, enumerative structures are good candidates for an annotation program; the frequent mixing of devices makes them an interesting case to test the functional equivalence between these different types of signaling; finally, their ability to occur at vastly different levels of text granularity is of particular interest in exploring the articulation between levels of text organization.

Within the annotation tool Glozz, topical chains were encoded as schemas consisting of a single unit with a set of topical expressions singled out that served to determine the extent of the segment, while enumerative structures were encoded as schemas composed of three different types of discourse units characterized respectively as trigger, items and closure and a set of units characterized as cues (e.g. sequencers, circumstances, connectives, parallelisms, etc.)

**Example 3.2.**

<table>
<tr><td>On the other hand, Saddam Hussein's regime has developed three important groups ..</td><td rowspan="3">TC</td><td rowspan="5">ES</td><td>TRIGGER</td></tr>
<tr><td>Though it reduced the Republican Guard by half, from 150000 to 70000 men, it made sure that the precious mechanised and armoured units were rebuilt. In order to do this it turned to illegal imports, but mostly it cannibalized equipment that had survived the bombing, often to the detriment of the army.</td><td>ITEM 1</td></tr>
<tr><td>The regime also moved away from a traditional air force toward a more operational air corps. It consolidated squadrons that were used to operate in close coordination with the Republican Guard.</td><td>ITEM 2</td></tr>
<tr><td>The importation of spare parts worked out to be easier for helicopters, which have the advantage of having a dual civilian and military status.<br>Finally, the almost daily incursions by American and British planes into the air exclusion zones, as well as the frequent attacks with cruise missiles, stimulated Saddam Husseins's interest in air defense units, renovated and pacified by privileges similar to those given to the Republican Guard. We stress that this is the main classical military move taken by Irak against a foreign adversary.</td><td></td><td>ITEM 3</td></tr>
<tr><td>To sum up, the regime has remodelled and redirected its armed forces in such a way as to move towards a more reliable and more compact system, both repressive and defensive in character.<br>In such a configuration, it no longer represents - despite the accusations coming from the USA - much of a menace for its neighbours.</td><td>TC</td><td>CLOSURE</td></tr>
</table>

Prior to annotation, a Biber-style systematic premarking of potentially relevant features (Biber, 1988) was automatically carried out on the POS-tagged and syntactically analyzed texts, with TreeTagger and SYNTEX (Bourigault, 2007). Premarked features, based on a wide range of studies of discourse markers, include visual devices and document structure signals such as headings, bulleted/numbered items (Power *et al.* , 2003), punctuation (e.g. paragraphs ending with [:], punctuational motifs such as [: ...; ...; and/or ...]), as well as lexico-syntactic features: coreferential and topical expressions (Cornish, 1999), item introducers (Hempel & Degand, 2008) ; prospective elements and anaphoric encapsulation (Francis, 1994) ; sentence-initial circumstancial adverbials – as potential frame introducers (Charolles M. *et al.* , 2005) – and other sentence-initial elements (e.g. connectives, appositions, etc.)

The human annotation proceeded in four steps. First, annotators detected ESs and TCs by scanning the text with the help of visual layout and highlighted premarked features. When a structure was detected, they indicated the boundaries of its subsegments: the topical chain segment for TCs, the trigger, items and closure for ESs. For

TCs, they identified all topical expressions by validating premarked features and adding new ones. For ESs, they indicated the expressions specifying the co-enumerability criterion (in boxes in example 3.2) and identified all features signalling the ES by validating premarked features and adding new ones. The step consisted in grouping sub-segments and features under a same schema. The annotation program began with a triple annotation of three texts by all three student annotators, with the option of consulting expert annotators in order to resolve problems with definitions and procedures. This led to an improved version of the manual. In a second stage, six texts were annotated by the three coders. The 27 annotated texts resulting from these two stages were used to measure inter-annotator agreement. Agreement was calculated in terms of F-measure, which gives an estimation of the average proportion of multi-level structures that two different coders have similarly identified in terms of text concerned, sub-segments for ESs and main referent for CTs. Results are 0.7 for ESs (i.e. 70% of ESs were conjointly annotated by two coders) and 0.65 for TCs. The 9 multiple annotated texts have since been post-annotated in order to produce a gold version. As the F-measures were deemed acceptable for this type of annotation, we proceeded with the last phase: annotation of 73 texts by one annotator per text.

As a whole, 1579 multi-level structures were annotated in 87 texts[5] (991 ESs and 588 TCs). Tables 3 give a quantitative overview of the results of the annotation campaign, in terms of the different objects presented above and for the three sub-corpora:

| corpus | ES | item | trigger | closure | TC | topical expr. |
|---|---|---|---|---|---|---|
| WIK2 | 401 | 1653 | 300 | 36 | 266 | 1853 |
| LING | 297 | 850 | 230 | 46 | 88 | 478 |
| GEOP | 293 | 863 | 209 | 49 | 234 | 1125 |
| Total | 991 | 3366 | 740 | 131 | 588 | 3456 |

Table 3: A quantitative overview of annotated Multi-level Structures (a)

As our discussion above of ANNODIS implies, from an analysis of inter annotator agreement, one can go two ways: one can either provide an expert reannotation as was done in the bottom up approach, or one can provide an adjudicated gold standard, as was done in the top down approach.

## 3.5 Annotation maintenance

The ANNODIS resource is available from REDAC (`http://redac.univ-tlse2.fr/corpus/`) under Creative Commons license BY-NC-SA 3.0 (Attribution - Non Commercial - Share Alike). For the bottom up approach, both the "naive" double annotations of the texts and the expert reannotations are available. Some post-processing was done before publishing it, and work in progress may lead us to publish new versions in the future. The post-processing mainly concerned annotation normalization (cues labelling for multi-level structures, rhetorical relation orientations) and annotation formatting for publishing. Work in progress includes qualitative analysis of annotated data, in order to refine or complete parts of the annotation.

[5] Taking into account the gold annotations rather than the annotations produced during the two first phases.

# 4 From annotated texts to applications and other linguistic forms

## 4.1 Linguistic Applications

The ANNODIS annotations have proved a useful resource on several fronts. The first explored was a validation of *the right frontier constraint* or RFC, a particular postulate of many discourse theories including SDRT. This work used the annotations from the bottom-up approach. The right frontier constraint (RFC) was originally proposed by (Polanyi, 1988) as a constraint on antecedents to anaphoric pronouns. Later, (Asher, 1993) adapted and refashioned this constraint in SDRT, postulating that an incoming discourse unit should attach either to the last discourse unit or to one that is superordinate to it via a series of subordinate relations and complex segments. Other discourse theories have similar constraints, though the empirical predictions of various versions of the RFC will depend on other assumptions made about discourse structure. Up until the study in (Afantenos & Asher, 2010), such postulates had never been validated empirically at a corpus level. They used the ANNODIS data from the "naive" phase in the bottom up annotation campaign in order to check the validity of SDRT's version of RFC. They found that the naive annotators, which had not been given any information on the structural postulates of SDRT, respected the RFC in 95% of the cases. The 5% remaining were mostly annotation errors due to the fact that the graphical tool used was not well adapted for this task. Besides being of interest to linguists and researchers on discourse structure, exploiting the RFC potentially has interesting computational implications: it can drastically reduce the search space for a discourse attachment, since we can consider as open to attachment only the nodes that are found on the RF.

The ANNODIS bottom-up annotations also proved valuable for research on discourse relations. Such studies help enrich discourse theories with an empirical basis. In our case, we have been able to use the corpus to provide SDRT with a better semantics for discourse relations and a better analysis of the cues triggering them. Most of the time, researchers use a *semasiological approach* to study discourse relations by looking at how various markers either trigger an inference to the presence of a discourse relation or block such an inference (Bras *et al.* , 2001; Bras, 2007; Bras *et al.* , 2009). Thanks to the discourse relation occurrences labelled in the ANNODIS corpus, *onomasiological approaches*, which start from the discourse relation annotation to discover various linguistic expressions associated with it, are possible. Such approaches help discover new markers for discourse relations, and are particularly interesting for discourse relations known to have few if any explicit discourse markers like Elaboration (Vergez-Couret, 2010). The annotation of Elaboration relations also showed bad inter-annotator agreement, which we explain by the existence of a multiplicity of cues that signal Elaboration. A qualitative analysis of the naive annotations of Elaboration corrected by Vergez-Couret helped expand the list of cues for Elaboration. (Atallah, 2014) examined the causal relations of the ANNODIS corpus and has refined the set of causal relations in SDRT. This work has shown that onomasiological approaches need much bigger corpora than the ANNODIS one and that markers of discourse relations mentioned in annotation manuals need to be as reliable as possible as cues; our annotation manual gave a table of linguistic markers, each associated to a list of possible discourse relations, which led to some wrong annotations with ambiguous markers, Finally, (Vergez-Couret, 2010) and (Atallah, 2014) showed that expert annotation is

essential for such linguistic research on discourse relations, which raises the question of the role of naive annotation.

The top down approach's study of enumerative structures, in particular their interaction with document structure (Ho-Dac *et al.* , 2010) and the combination of clues which signal them (Ho-Dac *et al.* , 2012), has also yielded interesting findings. ESs are an extremely frequent textual pattern, occurring in all sub-corpora, with a large diversity in size, textual granularity level, semantico-pragmatic function, with various forms of signalling. Data mining techniques show that ESs which interact explicitly with layout (e.g. ESs with subsection or bulleted/numbered items), tend to have a trigger which makes explicit the relation by which the enumerated items are related to each other. We are now examining the data from several qualitative angles in order to arrive at a functional characterisation of ESs, with a special interest for the link between particular forms of signalling and specific functions.

## 4.2 Computational applications

Discourse parsing is important and recognized to be a very difficult task in computational linguistics. The best methods to date incorporate some method of supervised machine learning over discourse annotations. Discourse parsing takes up the same three tasks that we outlined in section 2: text segmentation, attachment decisions, and the labeling of attachment arcs with discourse relations. ANNODIS provides us at least with a pilot test bed on which to test various proposals for discourse parsing. The ANNODIS resource has proved useful in developing automated methods for EDU segmentation.

Previous research on discourse segmentation has relied on the assumption that elementary discourse units (EDUs) in a document always form a linear sequence (i.e., they can never be nested). Unfortunately, this assumption turned out to be too strong for empirical reasons: given that parentheticals and appositions often have a scope out of local semantic operators, it makes sense to take them as separate discourse units, related typically to the clause or EDU that surrounds them by relations like E-elab, Commentary or Background. It thus proved fortunate that a theory like SDRT permitted such nesting. In (Afantenos *et al.* , 2010) we presented a simple approach to discourse segmentation that produced nested EDUs in the presence of appropriate environments. Our approach built on standard multi-class classification techniques combined with a simple repairing heuristic that enforces global coherence. Our system was developed and evaluated on the first round of annotations provided by the ANNODIS project. Cross-validated on only 47 documents (1,445 EDUs), our system achieved encouraging performance results with an F-score of 73% for finding EDUs.

We have also used the ANNODIS corpus for experiments on discourse parsing. Discourse parsing has to address the same questions about discourse structure that a theory or annotation manual does. Once EDUs have been identified, the next step in building a discourse structure for a text (or portion of text) is to determine the attachment of EDUs to other EDUs, the construction of larger CDUs and the labeling of the attachment links with a rhetorical relation. Most research in the area has focused on the task of relation labeling (Feng & Hirst, 2012) while discourse attachment has taken less attention by the community. Research on discourse structure also divides into two orthogonal categories: some researchers limit themselves to intra-sentential discourse structure (Sagae, 2009; Joty *et al.* , 2012); others tackle the problem of identifying the full discourse structure of a text (Hernault *et al.* , 2010; Subba & Di Eugenio, 2009).

The latter rely on "local" models to predict potential coherence relations, assuming independence between the decisions, and build the structure guided by greedy heuristics.

In (Muller *et al.* , 2012a) we proposed a more general approach to discourse structure prediction at the document level: (i) it performs a global search over the space of possible structures and optimizes a global criterion over the set of potential coherence relations; the global search is performed after estimating a probability distribution for attaching two arbitrary EDUs; (ii) a decoding mechanism is then applied, which can also take into account linguistically motivated constraints on the predicted structure. Specifically, our approach relies on the A* search algorithm, which is particularly well suited in allowing to capture constraints such as the Right Frontier Constraint.

We used maximum entropy- and Naive Bayes- based methods for the estimation of the local probability distributions and three different decoding mechanisms: i) a greedy one (essentially a reimplementation of (Hernault *et al.* , 2010)), ii) a maximum spanning tree approach (MST) on which no constraints can be encoded and iii) an A* decoder which can incorporate constraints, such as the RFC. Best results were achieved with MaxEnt and MST or A* (the difference had no statistical significance) and gave between 47 and 66% on the structure for the full set of relations and the reduced, 4-way classification. These results were difficult to align with discourse parsing experiments for inducing full discourse structures on text like those based on the RST tree bank (Hernault *et al.* , 2010; Subba & Di Eugenio, 2009), because of the different underlying structures used. However, Venant *et al.* (2013) shows that in fact these scores are comparable with results from larger corpora.

## 4.3 Opinion mining and preference extraction

Another area in which we have exploited the annotation model developed in the ANNODIS project was in the field of sentiment analysis. Sentiment analysis has become one of the most popular applications of natural language processing over the last decade both in academic research institutions and in companies. The goal of sentiment analysis is to extract automatically from a text an opinion held by the author or by agents described in the text about some object. One can do sentiment analysis either at the document (Turney, 2002) or the sentence level (Wiebe & Riloff, 2005).

Some of the authors of this paper participated in a recent project that used the Annnodis bottom-up annotation model, exploring the impact of discourse structure on the task of sentiment analysis.[6]on sentiment analysis with a study of French and English opinion texts.

Viewing opinions in a text as a simple aggregation of opinion expressions identified *locally* and hence taken in isolation is not appropriate, as shown in (4.1), an example extracted from our corpus of French movie reviews. (4.1) translated from the contains four opinions: the first three are strongly negative while the last one (introduced by the contrastive marker *but* in the last sentence) is positive. A bag of words approach would determines that this review is negative which is not the case here. Discourse structure provides a crucial between local and textual levels and hence is needed for a better understanding of the opinions expressed in texts (Asher *et al.* , 2008)(Somasundaran, 2010)(Trnavac & Taboada, 2010).

**Example 4.1.**

[6]The project was CASOAR, http://projetcasoar.wordpress.com, a two year DGA-RAPID project (2010-2012).

The characters are unsavory. The scenario is totally absurd. The decoration seems to be made out of cardboard. But all these elements make the charm of this TV series.

The data in the CASOAR project came from three corpora: (1) 181 French movie and product reviews (FMR) taken form AlloCine.fr for movie reviews, Amazon.fr for book and video game reviews and from Qype.fr for restaurant reviews, (2) 110 English movie reviews (EMR) from Metacritic and (3) 131 French news reactions (FNR) extracted from Lemonde.fr. The annotation scheme for CASOAR was multi-layered and included: (1) the expression level, (2) the opinion orientation of elementary discourse units and (3) the complete discourse structure according to the Segmented Representation Discourse Theory. Each level has its own annotation manual and annotation guide. The annotation scheme at the third level was inspired from the ANNODIS annotation manual that we modified by making explicit the structural constraints annotators should respect while building the discourse graph (such as the right frontier principle for example). When assuming that attachment is a yes/no decision on every EDUs pair, and that all decisions are independent, we obtained an F-measure of 69% for *FMR* and 68% for *FNR*. When commonly attached pairs were considered, we got a Cohen kappa of 0.57 for the full set of 17 relations for *FMR* and 0.56 for *FNR*. The results are a little bit higher compared to those obtained in the ANNODIS annotation campaign because the CASOAR annotation manual is more constrained and the corpora are smaller (an average of 20 EDUs compared to 55 EDUs in ANNODIS) which implies less long distance attachments.

In (Benamara *et al.* , 2015) it was shown that opinion and discourse structure are strongly related and that discourse is an important cue for sentiment analysis, at least for the corpus genre we have studied. The CASOAR corpus is a first step towards a discourse-based opinion analysis. We have already used a subset of this corpus (151 *FNR* documents, 1905 EDUs and 1766 discourse relations and 112 *FNR* documents, 835 EDUs and 924 relations) in order to investigate how discourse can help in the analysis of polarity (Chardon *et al.* , 2013b) and the assessment the overall opinion of a document (Chardon *et al.* , 2013a).

## 4.4 Further Annotation projects: Stac

Our ANNODIS and DISCOR annotation campaigns used texts. One might ask, does discourse annotation change substantially when moves to a different linguistic medium for imparting information, and if so how? In the project STrAtegic Conversation (STAC), we have begun to explore this question in an annotation campaign with a corpus of on-line chat dialogues involving negotiations in a popular board game that can be played on the internet. In contrast to ANNODIS, we have tried in this current annotation campaign to make our annotating instructions as explicit as possible with regards to structure as well as choice of relation and segmentation. Not surprisingly, the chat medium involves much shorter contributions, and turns become an important discourse segmentation device: from our initial experience here, it is rare that CDUs will span turns by more than one author and are often limited to a single turn. Turns have also made the segmentation process quicker, with the assumption that no EDU spans more than one turn. We have been able to automate significant parts of the segmentation process, requiring just an expert review of the machine given segmentation.

At the relational and structural level, differences between annotations on this corpus and the ANNODIS ones are more marked. First, an annotation campaign like ours has to decide how to handle relations between questions, assertions, and requests. In

| | |
|---|---|
| Continuation | Narration |
| Elaboration | Purpose |
| Conditional | Alternation |
| Explanation | Explanation* |
| Contrast | Correction |
| Result | Result* |
| Parallel | Clarification Q |
| Answer/ Question answer pair | Acknowledge |
| Q-elab/ follow up question | Commentary |

Table 4: Discourse Relations in STAC

this annotation campaign we have used many of the relations used in ANNODIS, but we needed to extend the relation set to handle relations involving questions. A natural and almost inescapable relation for dialogue annotation is one that involves some sort of answerhood relation between questions and their answers. However, we have noticed that relations like Elaboration can also hold between questions (Muller & Prévot, 2008), (Asher & Lascarides, 2003). The table below shows the current list of relations in use in the STAC annotation campaign.

The frequency of discourse relations in our dialogue corpus was quite different from the frequencies of discourse relations in text. The most frequent relations are Question Answer Pair, Q-elab (where a follow up question to typically another question asks for more details in order to provide an answer to the first question), Commentary and Acknowledgments. Elaborations and Explanations also are frequent. Elaborations typically occur, when an agent makes an offer and then further specifies it. This can often happen with questions:

**Example 4.2.**

A: Anyone want sheep for ore?
A: 2 sheep for 1 ore?

Acknowledgments, signaled by words like *OK, Right, Right then, Good, Fine, etc.* highlighted a challenge that we did not really address in ANNODIS (but see (Muller & Prévot, 2003), (Maudet *et al.* , 2006) for related discussion about acknowledgement scope). It's often difficult to determine whether the acknowledgment signals an understanding of what was said, an acceptance of what was said or an acceptance and a signal to change the topic of conversation or move on. It's also often difficult to determine what is being acknowledged. The difficulty in determining the scope of a discourse relation is a general one, but with acknowledgments it was especially obvious. To handle these challenges, we have allowed the annotators to leave this last feature partially specified or unspecified.

# 5 Conclusions

We've given in this chapter an overview of our efforts over the past decade to find good annotation models for discourse structures in texts and dialogues. Annotating discourse structure on constructed examples is a challenging task; annotating real texts, be they

monologues or dialogues, well is even harder. Part of the reason is that we still don't have a robust and detailed theoretical grasp of what discourse structure is nor how such structures are conveyed in language. But in order to progress in our theoretical understanding, we need to look at more data; and so annotation efforts and theoretical understanding are really of a piece, each feeding the other and each needing the other in successive rounds of a dialectic.

# References


Afantenos, Stergos, Asher, Nicholas, Benamara, Farah, Bras, Myriam, Fabre, Cécile, Ho-Dac, Lydia-Mai, Le Draoulec, Anne, Muller, Philippe, Péry-Woodley, Marie-Paule, Prévot, Laurent, Rebeyrolle, Josette, Tanguy, Ludovic, Vergez-Couret, Marianne, & Vieu, Laure. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. *In:* Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, & Piperidis, Stelios (eds), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Afantenos, Stergos D., & Asher, Nicholas. 2010. Testing SDRT's Right Frontier. *Pages 1–9 of: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*.

Afantenos, Stergos D., Denis, Pascal, Muller, Philippe, & Danlos, Laurence. 2010. Learning Recursive Segments for Discourse Parsing. *In: Proceedings of LREC 2010*.

Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Kluwer.

Asher, N., Hardt, D., & Busquets, J. 2001. Discourse Parallelism, Ellipsis and Ambiguity. *Journal of Semantics*, **18**(1).

Asher, Nicholas. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press.

Asher, Nicholas, & Lascarides, Alex. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

Asher, Nicholas, Benamara, Farah, & Mathieu, Yvette Yannick. 2008. Distilling Opinion in Discourse: A Preliminary Study. *Pages 7–10 of: Proceedings of Computational Linguistics (CoLing)*.

Atallah, Caroline. 2014. *Analyse de relations de discours causales en corpus : étude empirique et caractérisation théorique*. Ph.D. thesis, Université de Toulouse, Toulouse.

Baldridge, J., Asher, N., & Hunter, J. 2007. Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts. *Zeitschrift fur Sprachwissenschaft*, **26**, 213–239.

Benamara, Farah, Asher, Nicholas, Mathieu, Yannick, Popescu, Vladimir, & Chardon, Baptiste. 2015. Evaluation in Discourse: a Corpus-Based Study. *Dialogue and Discourse*. in press.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Bourigault, D. 2007. *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'HDR, Université de Toulouse.

Bras, Myriam. 2007. French adverb *d'abord* and Discourse Structure. *Pages 77–102 of:* Aurnague, Michel, Larrazabal, Jesus-Mari, & Korta, Kepa (eds), *Language, Representation and Reasoning. Memorial Volume to Isabel Gomez Txurruka*. Bilbao: Presses Universitaires du Pays Basque.

Bras, Myriam, Le Draoulec, Anne, & Vieu, Laure. 2001. French Adverbial *Puis* between Temporal Structure and Discourse Structure. *Pages 109–146 of:* Bras, Myriam, & Vieu, Laure (eds), *Semantic and Pragmatic Issues in Dialogue: Experimenting with Current Theories*. CRISPI, vol. 9. Amsterdam: Elsevier.

Bras, Myriam, Le Draoulec, Anne, & Asher, Nicholas. 2009. A formal analysis of the French Temporal Connective *alors*. *Oslo Studies in Language*, **1**, 149–170.

Carletta, J., Isard, S., & Doherty-Sneddon, G. 1996. *HCRC dialogue structure coding manual*. HCRC Publications, The University of Edinburgh.

Chafe, W.L. 1994. *Discourse Consciousness and Time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press: Chicago.

Chardon, Baptiste, Benamara, Farah, Mathieu, Yvette Yannick, Popescu, Vladimir, & Asher, Nicholas. 2013a. Measuring the Effect of Discourse Structure on Sentiment Analysis. *Pages 25–37 of: CICLing*.

Chardon, Baptiste, Benamara, Farah, Mathieu Yannick, Yvette, Popescu, Vladimir, & Asher, Nicholas. 2013b. Sentiment Composition Using a Parabolic Model. *Pages 47–58 of: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Charolles, Michel. 1997. L'encadrement du discours - Univers, champs, domaines et espace. *Cahiers de recherche linguistique*, **6**, 1–73.

Charolles M., Le Draoulec A., Péry-Woodley, M.-P., & Sarda, L. 2005. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, **15**(2), 203–218.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Colléter, M., Fabre, C., Ho-Dac, L.-M., Péry-Woodley, M.-P., Rebeyrolle, J., & Tanguy, L. 2012. *La ressource ANNODIS multi-échelle : guide d'annotation et bonus*. Tech. rept. 20. Carnets de grammaires, CLLE-ERSS.

Cornish, F. 1999. *Anaphora, Discourse and Understanding. Evidence from English and French.* Clarendon Press: Oxford.

Danlos, Laurence. 2008. Strong generative capacity of RST, SDRT and discourse dependency DAGSs. *Pages 69–95 of:* Benz, A., & Kuhnlein, P. (eds), *Constraints in Discourse*. John Benjamins.

Egg, Markus, & Redeker, Gisela. 2010. How Complex is Discourse Structure? *In:* Calzolari, N., Choucri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. (eds), *Proceedings of LREC'10*. ELRA.

Enkvist, N.E. 1989. Connexity, Interpretability, Universes of Discourse, and Text Worlds. *Pages 162–186 of:* Allén, J. (ed), *Possible Worlds in Humanities, Arts and Sciences*. Walter de Gruyter: Berlin/New-York.

Feng, Vanessa Wei, & Hirst, Graeme. 2012. Text-level Discourse Parsing with Rich Linguistic Features. *Pages 60–68 of: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics.

Forbes, Katherine, Miltsakaki, Eleni, Prasad, Rashmi, Sarkar, Anoop, Joshi, Aravind K., & Webber, Bonnie L. 2003. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information*, **12**(3), 261–279.

Francis, Gill. 1994. *Labelling Discourse: An Aspect of Nominal-Group Lexical Cohesion*. London-New York: Routledge. Pages 83–101.

Fries, P. 1995. Themes Method of Development and texts. *Pages 317–359 of:* R.Hasan, & P.Fries (eds), *On Subject and Theme: A Discourse Functional Perspective*. John Benjamins: Amsterdam/Philadelphia.

Goutsos, Dyonisos. 1996. A model of sequential relations in expository test. *Text*, **16**(4), 501–533.

Grosz, B., & Sidner, C. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, **12**, 175–204.

Halliday, M.A.K. 1977. Text as semantic choice in social contexts. *Pages 176–226 of:* van Dijk, T., & Petöfi, J.S. (eds), *Grammars and Descriptions*. Walter de Gruyter: Berlin.

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar*. 2nd edn. London: Arnold.

Halliday, M.A.K., & Hasan, R. 1976. *Cohesion in English*. Longman: London.

Hempel, Susanne, & Degand, Liesbeth. 2008. Sequencers in Different Text Genres: Academic Writing, Journalese and Fiction. *Journal of Pragmatics*, **40**, 676–693.

Hernault, Hugo, Prendinger, Helmut, duVerle, David A., & Ishizuka, Mitsuru. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, **1**(3), 1–33.

Hitzeman, J., Moens, M., & Grover, C. 1995. Algorithms for Analyzing the Temporal Structure of Discourse. *Pages 253–260 of: Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics*.

Ho-Dac, L.-M., & Péry-Woodley, M.-P. 2009. A data-driven study of temporal adverbials as discourse segmentation markers. *Discours*, **4**.

Ho-Dac, L.-M., Péry-Woodley, M.-P., & Tanguy, L. 2010. Anatomie des structures énumératives. *In: Actes de TALN 2010*. Montréal: Université de Montréal, for ATALA.

Ho-Dac, L.-M., Fabre, C., Péry-Woodley, M.-P., Rebeyrolle, J., & Tanguy, L. 2012. On the signalling of multi-level discourse structures. *Discours*, **10**.

Hobbs, J.R.ì979. Coherence and Coreference. *Cognitive Science*, **3**(1), 67–90.

Hovy, Eduard H. 1990 (June). Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. *Pages 128–136 of: Proceedings of the Fifth International Workshop on Natural Language Generation*.

Joty, Shafiq, Carenini, Giuseppe, & Ng, Raymond. 2012. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. *Pages 904–915 of: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics.

Kamp, H., & Reyle, U. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Landis, JR, & Koch, GG. 1977. The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

Lascarides, A., & Asher, N. 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, **16**(5), 437–493.

Mann, W., & Thompson, S. 1987. *Rhetorical Structure Theory : a theory of text organization.* Tech. rept. Information Science Institute.

Marcu, Daniel. 1996. Building up rhetorical structure trees. *Pages 1069–1074 of: Proceedings of the thirteenth national conference on Artificial intelligence - Volume 2*. AAAI'96. AAAI Press.

Mathet, Y., & Widlöcher, A. 2009. La plate–forme GLOZZ : environnement d'annotation et d'exploration de corpus. *In: Actes de TALN 2009*. Senlis: LIPN, for ATALA.

Maudet, Nicolas, Muller, Philippe, & Prévot, Laurent. 2006. Social Constraints on Rhetorical Relations in Dialogue. *Pages 133–139 of:* Sidner, Candy, Harpur, John, Benz, Anton, & Kühnlein, Peter (eds), *Proceedings of the Workshop on Constraints in Discourse*.

Muller, P., & Prévot, L. 2003 (Sept 4th-6th). An empirical study of acknowledgment structures. *In: Proceedings of Diabruck 2003, 7th workshop on the semantics and pragmatics of dialogue*.

Muller, P., Afantenos, S., P., Denis, & Asher, N. 2012a. Constrained decoding for text-level discourse parsing. *In: Proceedings of COLING*.

Muller, P., Vergez, M., Prevot, L., Asher, N., Benamara, F., Bras, M., Le Draoulec, A., & Vieu, L. 2012b (december). *Manuel d'annotation en relations de discours du projet Annodis*. Tech. rept. 21. CLLE.

Muller, Philippe, & Prévot, Laurent. 2008. The rhetorical attachment of questions and answers. *In:* Korta, Kepa, & Garmendia, Joana (eds), *Meaning, Intentions, and Argumentation*. (CSLI-LN) Center for the Study of Language and Information - Lecture Notes, vol. 186. http://www.journals.uchicago.edu/: University of Chicago Press.

Polanyi, Livia. 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, **12**, 601–638.

Polanyi, Livia, Culy, Chris, van den Berg, Martin, Thione, Gian Lorenzo, & Ahn, David. 2004. A Rule Based Approach to Discourse Parsing. *Pages 108–117 of:* Strube, Michael, & Sidner, Candy (eds), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*. Cambridge, Massachusetts, USA: Association for Computational Linguistics.

Power, R., Scott, D., & Bouayad-Agha, N. 2003. Document Structure. *Computational Linguistics*, **2**(29), 211–260.

Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind, & Webber, Bonnie. 2008. The Penn Discourse TreeBank 2.0. *In:* Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odjik, Jan, Piperidis, Stelios, & Tapias, Daniel (eds), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Prévot, Laurent, Asher, Nicholas, & Vieu, Laure. 2009. Une formalisation plus précise pour une annotation moins confuse: la relation d'Élaboration d'entité. *Journal of French Language Studies*, **19**(2), 207–228.

Roze, Charlotte. 2013. *Vers une algèbre des relations de discours*. Ph.D. thesis, Université Paris 7.

Sagae, Kenji. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. *Pages 81–84 of: Proceedings of the 11th International Conference on Parsing Technologies*. IWPT '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Somasundaran, Swapna. 2010. *Discourse-level relations for Opinion Analysis*. Ph.D. thesis, University of Pittsburgh.

Subba, Rajen, & Di Eugenio, Barbara. 2009. An effective Discourse Parser that uses Rich Linguistic Information. *Pages 566–574 of: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics.

Trnavac, R., & Taboada, M. 2010. The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, **34(3)**, 301–318.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *In: Proceedings of Annual Meeting of the Association for Computational Linguistics*.

Venant, Antoine, Asher, Nicholas, Muller, Philippe, Denis, Pascal, & Afantenos, Stergos. 2013. Expressivity and comparison of models of discourse structure. *Pages 2–11 of: Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics.

Vergez-Couret, Marianne. 2010. *Etude en corpus des réalisations linguistiques de la relation d'Elaboration*. Ph.D. thesis, Université de Toulouse, Toulouse.

Webber, Bonnie, Egg, Markus, & Kordoni, Valia. 2012. Discourse structure and language technology. *Natural Language Engineering*, **18**(4), 437–490.

Wiebe, Janyce, & Riloff, Ellen. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *Pages 486–497 of: Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. Lecture Notes in Computer Science, vol. 3406.

Wolf, Florian, & Gibson, Edward. 2005. Representing Discourse Coherence: A Corpus Based Study. *Computational Linguistics*, **31**(2), 249–287.