An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus

Stergos Afantenos\textsuperscript{1}, Nicholas Asher\textsuperscript{1}, Farah Benamara\textsuperscript{1}, Myriam Bras\textsuperscript{2}, Cécile Fabre\textsuperscript{2}, Mai Ho-dac\textsuperscript{2}, Anne Le Draoulec\textsuperscript{2}, Philippe Muller\textsuperscript{1}, Marie-Paule Péry-Woodley\textsuperscript{2}, Laurent Prévot\textsuperscript{2}, Josette Rebeyrrolle\textsuperscript{2}, Ludovic Tanguy\textsuperscript{2}, Marianne Vergez-Couret\textsuperscript{2}, Laure Vieu\textsuperscript{1}

\textsuperscript{1}IRIT, Univ. Toulouse, France
\textsuperscript{2}CLLE, Univ. Toulouse, France
\textsuperscript{3}LPL, Univ. de Provence, France
(authors are in alphabetical order)

Abstract
We describe the Annodis corpus of discourse structures for French. The corpus joins two perspectives on discourse on a variety of textual genres: a bottom-up approach and a top-down approach. The bottom-up view builds incrementally a structure from elementary discourse units, while the top-down view focuses on the selective annotation of multi-level discourse structures. The corpus is composed of texts that are diversified with respect to genre, length and type of discursive organisation. The methodology we followed involved an iterative design of annotation guidelines in order to reach satisfactory inter-annotator agreement levels. This allows us to raise a few issues relevant for the comparison of such complex objects as discourse structures. The corpus is seen as a source of empirical evidence for discourse theories, and we present the first analyses using the annotations: for instance testing hypotheses on constraints governing the structure of discourse, studying special constructs such as enumerations.

Keywords: Corpus linguistics, discourse analysis

1. Introduction

This paper describes the ANNODIS resource, a diversified corpus of written French texts enriched with several kinds of markup, including a manual annotation of discourse structures. The manual annotation is based on two approaches to discourse: a “bottom-up” approach whose aim is to construct the structure of a discourse from elementary units linked by coherence relations, and a “top-down” or “macro” approach which focuses on the selective annotation of multi-level discourse structures.

The ANNODIS corpus is the first such resource in French to our knowledge. But it also has distinct characteristics in comparison with English discourse annotated corpora like the Penn Discourse TreeBank or the RST tree bank. It is composed of texts that are diversified with respect to genre, length and type of discursive organisation. It contains two distinct and complementary types of annotation. The bottom-up approach aims to provide a complete discourse structure for each text, starting from a segmentation of the text into elementary discourse units (EDUs), and then linking these by means of discourse relations, also known as coherence or rhetorical relations, to form complex discourse units or CDUs, which in turn may be linked via discourse relations to other discourse units. The bottom-up approach specifies completely the semantic scope of each discourse relation, making transparent an interpretation of the text that takes into account the semantic effects of discourse relations.

The top-down or “macro” approach focuses on two text-organizing strategies realized at different levels of textual granularity (from less than a paragraph to several sections), hence “multi-level” discourse structures: enumerative structures and topical chains. The bottom-up approach exploits cues based on syntax, discourse markers and deep semantics, while the top-down approach stresses the role of document structure (the text’s graphical constituents) in its interaction with other cues. Combining both of these annotation approaches together creates the potential for some novel synergies. The top-down approach provides a macro level organization that constrains the construction of CDUs in the bottom-up approach, while the bottom-up approach provides detailed decompositions and semantic analysis of the structures that the top-down approach takes as primitive.

2. The Annotation Campaign

2.1. Corpus and Annotation Interface

The Annodis corpus is in two parts, each corresponding to an approach and annotation scheme. The bottom-up corpus consists of short texts (a few hundred words each) as the annotation process aimed at a detailed analysis of every discourse unit. This annotation method can also target excerpts from longer documents. The top-down approach, on the other hand, with its annotation scheme focusing on high-level discourse structures occurring at different levels of textual granularity, required longer more complex documents (several thousands words each).

In order to provide a diversified corpus, we selected texts that show variations along three parameters: genre, type and document structure. Four text genres are represented in the corpus, originating from different sources: short news articles from the French daily \textit{Est Républicain}, encyclopedia articles (from the French Wikipedia), linguistics research papers (from \textit{Colloque Mondial de Linguistique française}) and international relation reports (from the \textit{Institut Français de Relations Internationales}). These sources each favour a dominant text-type: narrative, expository, or argumentative. Finally, document structure is a rough measure of the amount of structuring features found in the doc-
2.2. Bottom-Up Approach

The bottom-up approach used both naive and expert annotators. We have performed three phases of annotation. During the first preliminary phase two graduate-level students annotated 50 documents. We used their input in order to create an annotation manual which was used afterwards during the second, so called, “naive” phase. During this second phase 3 undergraduate students with no knowledge whatsoever of discourse theories doubly annotated 86 documents. The annotators were trained for a week, with the help of the aforementioned manual and the graphical annotation tool Glozz, designed to help them segment and annotate the documents as described in the previous section.

It is important at this point to further expand on the nature of the provided manual and the information that was transmitted via it to the naive annotators. The manual focused essentially on two different aspects of the discourse annotation process: segmentation and typology of relations. Concerning the first, annotators were provided 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. The manual then went on to describe the discourse relations that could link two discourse units. The goal of the manual was the development of an intuition for each relation, suitable for the level of the annotators. Occasional examples were provided, but we tried to avoid exhaustively listing the possible discourse markers that could trigger any particular relation.

Crucially, the manual did not provide any details of the structural postulates of the underlying theory. More specifically we did not mention anything on distance of attachment, crossed dependencies and more theoretical postulates, such as contraints on attachments (the so-called “right frontier” of discourse structure, see section 3.1.2.), mainly because we wanted to test the intuitions of the naive annotators. 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 complex discourse unit (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 figure 1.

![Figure 1: An example of discourse graph.](image)

During the third and last phase, expert annotators adjudicated the naive annotation on the 86 documents and corrected them.

The view of discourse structure underlying our approach is that common to the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), the Linguistic Discourse Model (LDM) (Polanyi et al., 2004) the GraphBank model (Wolf and Gibson, 2005), the Discourse Lexicalized Tree Adjoining Grammar model (DLTAG) (Forbes et al., 2003), the Penn Discourse Treebank model (PDTB) (Prasad et al., 2008), and the Segmented Discourse Representation Theory (SDRT) (Asher, 1993). Most of these theories define hierarchical structures by constructing CDUs from EDUs in recursive fashion. In RST, for example, discourse is represented as a rooted tree in which adjacent EDUs are grouped together into complex discourse units which are then recurrently connected with other adjacent elementary or complex discourse units (called spans in the RST jargon). Depending on the relation linking two spans, the spans can serve either as nuclei or as satellites to the relation, satel-
lites being more important for the relation. In contrast to RST, PDTB does not focus on structure at all, but just on discourse relations and the explicit or implicit discourse markers that can trigger those relations. A common ground between RST and PDTB is that they both focus on adjacent discourse units in order to assign a discourse relation to that pair. The GraphBank model (as well as SDRT), on the other hand, go beyond adjacent discourse units allowing for the creation of full discourse graphs which capture complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups, as well as crossed dependencies, etc.

In our case SDRT served as the point of departure for the bottom-up annotation, as it provides a graph-based view of discourse structure, which is arguably more expressive than that of other theories (Danlos, 2007). The bottom-up approach focuses on providing a complete structure of a text, starting from the segmentation into EDUs (mostly clauses, appositions, some adverbials). Semantically, each EDU contains at least one eventuality description, and often only one. The relations linking DUs in this approach are a set of relations that were chosen because they are more or less common to all the theories of discourse mentioned above, or correspond to well-defined subgroups in fine-grained theories. The intermediate level of granularity was chosen as a compromise between informativeness and reliability of the annotation process. It corresponds to the level chosen in PDTB, and a coarse-grained RST. We used earlier work on these relations and how they are linguistically marked to guide the annotation process. The linguistic cues include not only so-called discourse markers but also tense and aspectual shifts, as well as specific syntactic structures. The list of relations used is the following: EXPLANATION, GOAL, RESULT, PARALLEL, CONTRAST, CONTINUATION, ALTERNATION, ATTRIBUTE, BACKGROUND, FLASHBACK, FRAME, TEMPORAL-LOCATION, ELABORATION, ENTITY-ELABORATION, COMMENT. Most of these are self-explanatory. Less obvious may be “Frame”, which relates a framing adverbial and EDUs within its scope (e.g. During the 20th century, [EDU1], [EDU2] etc.), and “Entity-elaboration”, which relates an EDU and a clause pertaining to an object within that clause (e.g. President X, who resigned yesterday, declared that Y.)

Table 2 shows the number of discourse units that were annotated in the corpus, and a breakdown by sub-corpus. We distinguish elementary discourse units and explicit complex units. Table 3 shows a breakdown of the relation types found in the corpus.

<table>
<thead>
<tr>
<th>relation type</th>
<th>corpus total</th>
<th>Est Rep</th>
<th>Wikip.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb Texts</td>
<td>87</td>
<td>39</td>
<td>42</td>
</tr>
<tr>
<td>Nb words</td>
<td>28146</td>
<td>9768</td>
<td>17330</td>
</tr>
<tr>
<td>EDU</td>
<td>3188</td>
<td>1159</td>
<td>1949</td>
</tr>
<tr>
<td>CDU</td>
<td>1395</td>
<td>510</td>
<td>829</td>
</tr>
</tbody>
</table>

Table 2: Discourse unit counts in expert annotations

2.3. Multi-level Structures annotation in a top-down approach

The concern of the top-down approach is with text organisation strategies, viewed in a Systemic Functional framework, and in particular with strategies regarding textual continuity and discontinuity (Goutsos, 1996; Martin, 2001). To translate this broad concern into a realistic annotation programme, we focused on two multi-level discourse structures (i.e. from two sentences up to several headed sections): topical chains and enumerative structures. Topical chains (TCs) are a specific type of cohesive chain (Halliday and 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 as illustrated in the example in Fig 2.

Enumerative structures (ESs) are segments resulting from a text organisation strategy which consists in packaging together independent elements according to a specific interpretation criterion. (Luc et al., 2000) give the following definition:

"The textual act [of Enumerating] consists in textually transposing the co-enumerability of the listed entities into the co-enumerability of the linguistic segments describing them, which thereby become the entities constituting the enumeration (the items). The identity of status of the items in the enumeration expresses the identity of status of the listed entities in the world". (Luc et al., 2000, p 25, our translation).

On either side of the enumeration proper, two optional segments may be found: a trigger and a closure. Enumera-
La LAF, rédigée en collaboration avec Igor Mel'cuk, est un ouvrage qui a déjà mentionné à la section 4.1. En tant qu’ouvrage publié, il tire son originalité du fait qu’il est à la fois un manuel de lexicologie des-\textit{tine}, en tout premier lieu, aux enseignants de langue et un \textit{échantillon} de dictionnaire du français, reposant sur une adaptation des descriptions formalisées de la LEC. Il s’accompagne d’un site web, où sont notamment rendus disponibles pour les enseignants de français des modèles d’exercices visant l’apprentissage du vocabulaire. Par sa finalité et par sa double nature (présentation de notions lexicologiques et de descriptions lexico-syntactiques), la LAF peut être rapprochée de Picche (2007). Il est intéressant de constater que le travail d’interfacage des principes et descriptions de la LEC opéré lors de la rédaction du LAF a permis, de façon rétroactive, de faire progresser l’approche théorique elle-même. On trouvera un bilan de l’expérience acquise au cours de la rédaction du LAF dans Polguère (2007). Dans ce texte, on fait notamment état des innovations introduites pour ce qui est de la caractérisation sémantique des unités lexicalas (au moyen d’étiquettes sémantiques) et de l’encodage des relations lexicales paradigmatiques et syntagmatiques (au moyen de formules dites « de vulgarisation »). Une autre caractéristique originale du LAF est sa méthodologie d’élaboration (Polguère, 2000b). Il est en effet entièrement dérivé de la base lexicale DiCo des dérivations sémantiques et collocations du français, développée par Igor Mel’cuk et le présent auteur. Cette façon de procéder assure au LAF une rigueur formelle sous-jacente et, surtout, nous permet de s’assurer de la base source DiCo d’autres « produits », comme celui dont il va maintenant être question.

Le LAF. \textit{Des orientations d’action}

<table>
<thead>
<tr>
<th>TC</th>
<th>ITEM 1</th>
<th>ITEM 2</th>
<th>ITEM 3</th>
<th>ITEM 4</th>
<th>CLOSURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITEM 1</td>
<td>- Mieux organiser notre politique étrangère dans la région ce qui passe, notamment, par la mise en place de structures permettant […]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM 2</td>
<td>- Accentuer notre coopération avec des partenaires d’influence, notamment en établissant une coopération renforcée avec certains […]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM 3</td>
<td>- Manifester notre souci de voir émerger des systèmes démocratiques dans la région en développant une politique d’influence auprès des […]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITEM 4</td>
<td>- Contribuer plus efficacement à la solution des principales crises régionales, ce qui comporterait les actions suivantes: […]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLOSURE</td>
<td>En conclusion, les turbulences qui affectent le moyen orien-tent ont atteint un niveau de haute intensité qui représente, pour les pays occidentaux et, plus spécialement, pour l’Europe, de grands risques, notamment […]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: ES – Enumerative Structure – covering a whole subsection: first, the heading together with the opening paragraph announce that the following text will list four themes of directions for action (re. the relationship between France and the Middle East); next, four bulleted items detail each of these themes, which are thereby presented as co-enumerable, i.e. identical in status with regard to the co-enumerability criterion; finally, the last paragraph of the subsection closes the enumeration with a conclusion.

Table 3: Discourse relations of the expert annotations

<table>
<thead>
<tr>
<th></th>
<th>total (Nb)</th>
<th>(%)</th>
<th>ER</th>
<th>(%)</th>
<th>WK</th>
<th>(%)</th>
<th>total (Nb)</th>
<th>(%)</th>
<th>ER</th>
<th>(%)</th>
<th>WK</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternation</td>
<td>18</td>
<td>0.5</td>
<td>0.3</td>
<td>0.6</td>
<td>explanation</td>
<td>130</td>
<td>3.9</td>
<td>4.4</td>
<td>3.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>attribution</td>
<td>75</td>
<td>2.2</td>
<td>3.0</td>
<td>1.7</td>
<td>feedback</td>
<td>27</td>
<td>0.8</td>
<td>1.4</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>background</td>
<td>155</td>
<td>4.6</td>
<td>5.2</td>
<td>4.8</td>
<td>frame</td>
<td>211</td>
<td>6.3</td>
<td>6.2</td>
<td>5.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comment</td>
<td>78</td>
<td>2.3</td>
<td>3.6</td>
<td>1.3</td>
<td>goal</td>
<td>95</td>
<td>2.8</td>
<td>3.1</td>
<td>2.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>continuation</td>
<td>681</td>
<td>20.3</td>
<td>20.1</td>
<td>21.1</td>
<td>narration</td>
<td>349</td>
<td>10.4</td>
<td>11.1</td>
<td>10.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contrast</td>
<td>144</td>
<td>4.3</td>
<td>3.7</td>
<td>4.6</td>
<td>parallel</td>
<td>59</td>
<td>1.8</td>
<td>2.2</td>
<td>1.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eelab</td>
<td>527</td>
<td>15.7</td>
<td>14.1</td>
<td>16.4</td>
<td>result</td>
<td>163</td>
<td>4.9</td>
<td>4.7</td>
<td>5.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>elaboration</td>
<td>625</td>
<td>18.6</td>
<td>16.3</td>
<td>19.4</td>
<td>temploc</td>
<td>18</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: TC – Topical Chain – covering two paragraphs. All sentences contain topical expressions referring to le LAF except for the sentence in italics. Topical expressions are in bold.

Table 2: Discourse relations of the expert annotations

Figure 1: LE – Lexico-semantic Expressions – covering a section

Figure 2: TC – Topical Chain – covering two paragraphs. All sentences contain topical expressions referring to le LAF except for the sentence in italics. Topical expressions are in bold.

tative structures are therefore characterised by an internal organisation involving three sub-segments: an optional trigger announcing the enumeration; several items composing the enumeration (at least two items must be identified for a structure to be present); an optional closure which summarises 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 given in Fig 3, "thèmes" is such an expression. We call such lexical expressions enumeraTheme.

The annotation method for these two multi-level structures, fully described in the annotation manual, distinguishes two stages: (1) identifying multi-level structures and delimiting segments (TCs and ESs) and sub-segments (triggers, items, closures); (2) identifying the features signalling these structures (topical cues and trigger/item/closure cues).

Prior to annotation, a Biber-style systematic pre-marking of potentially relevant features (Biber, 1988; Biber et al., 2007) was automatically carried out on the POS-tagged and syntactically analysed texts (with TreeTagger and SYN-TEX (Bourigault, 2007)). Visualisation of this pre-marking was used during the annotation process in order to help annotators identify the structures and the features signalling them. Pre-marked 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; Péry-Woodley and Scott, 2006), punctuation (e.g. paragraphs ending with [...]; punctuational motifs such as [...]; ...; ...; and/or [...]), as well as lexicosyntactic features. Via specifically-built lexica and local grammars, the following lexico-syntactic features were pre-marked: coreferential and topical expressions (Cornish, 1999; Grosz et al., 1995; Gundel, 1998) e.g. pronouns and lexical reiterations; item introducers (Turco and Coltier, 1988; Jackiewicz, 2005; Hempel and Degand, 2008) e.g. firstly, finally, the first X, on the other hand, ; predictive elements and anaphoric encapsulation (Francis, 1994; Bras et al., 2008; Legallois, 2006) ; sentence-initial circumstan-tial adverbials (as potential frame introducers (Charolles,
notations produced during the two first phases.

The manual annotation process was as follows: the text was loaded into the Glozzz interface and coders went on to detect ESs and TCs by scanning the text with the help of visual layout and highlighted pre-marked features. For each structure detected, the coders indicated the boundaries of its segments and sub-segments, and, in the case of ESs, the enumaTheme, i.e. the expression specifying the co-enumerability criterion. Finally, they annotated features signalling these (sub-)segments by validating pre-marked features seen as relevant, and by identifying additional features that had not been pre-marked (such as syntactic parallelism, trigger reiteration).

The annotation was organized in three stages. Initially, three texts were annotated by all three coders, 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 the 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 in terms of F-measure, with results of 0.7 for ESs and 0.65 for TCs (calculated by comparing annotated boundaries and cues). These 27 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, 1316 multi-level structures were annotated in 82 texts2 (829 ESs and 487 TCs). Tables 4 and 5 give a quantitative overview of the results of the annotation campaign, in terms of the different objects presented above and the different sub-corpus:

<table>
<thead>
<tr>
<th>corpus</th>
<th>added features</th>
<th>validated premarked features</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIK2</td>
<td>1677</td>
<td>2428</td>
</tr>
<tr>
<td>LING</td>
<td>937</td>
<td>708</td>
</tr>
<tr>
<td>GEOP</td>
<td>1130</td>
<td>993</td>
</tr>
<tr>
<td>ANNODIS</td>
<td>3744</td>
<td>4129</td>
</tr>
</tbody>
</table>

Table 5: A quantitative overview of annotated Multi-level Structures (b)

### 3. First Experiments and Analyses

#### 3.1. Rhetorical Relations

The availability of annotated discourses should allow for a number of empirical studies on pragmatic phenomenon. It can also feed work on automated predictions of discourse structures. We present here the efforts that are under way within the project, which already yielded some interesting results.

##### 3.1.1. EDU segmentation

EDU segmentation is the task of automatically finding the boundaries of elementary units of discourse structure in a text. This is the first stage of a full structure prediction. We cast the task of EDU identification as a classification problem on the level of tokens. More specifically, each token can either start or end an EDU, be an EDU by itself, or be strictly contained within an EDU.3

We built a four-class classifier that maps each token \( w_i \) in a discourse \( w_1, \ldots, w_n \) to one of the following boundary types \( B = \{ \text{left}, \text{right}, \text{both}, \text{nothing} \} \). These correspond to the different bracketing configurations found in our corpus, respectively (i) \( w_i \) opens a segment, (ii) \( w_i \) ends a segment, (iii) \( w_i \) is a single-token segment, and (iv) none of the above.

For our classifier, we used a regularized maximum entropy (MaxEnt, for short) model (Berger et al., 1996).

The classification was followed by a post-processing enforcing well-balanced segments. After post-processing we had an F-measure of 0.735 for the EDUs as a whole. We present more details in (Afantenos et al., 2010).

##### 3.1.2. Determining attachment points and the right frontier constraint

The right frontier constraint (RFC) was originally proposed by (Polanyi, 1988) as a constraint on antecedents to anaphoric pronouns. Later, (Asher, 1993) took this constraint in SDRT postulating that an incoming discourse unit should attach either to the last discourse unit or to one that is super-ordinate to it via a series of subordinate relations and complex segments. This postulate was never validated empirically at a corpus level. We used the Annodis data from the “naive” phase in order to check its validity. We found that the naive annotators, which had not been given any information on the structural postulates of SDRT, have respected the RFC in 95% of the cases. The 5% remaining was mostly annotation errors due to the fact that the graphical tool used was not well adapted for this task. More details are in (Afantenos and Asher, 2010). One practical implication is that the RFC 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.

##### 3.1.3. Evaluating agreement on complex relational data

Evaluating agreement on complex relational data such as discourse annotations is far from obvious, and collecting this corpus has raised a number of interesting issues from this perspective. Two kinds of information are annotated with a discourse graph: the attachment of discourse units to each other, and the labelling of the attachment arcs via discourse relations. We thus have two types of agreement to define, and the second one (relations labels) depend on the agreement for the first one (discourse unit pairs). We leave

2Taking into account the gold annotations rather than the annotations produced during the two first phases.

3In contrast to other theories EDUs in SDRT can be embedded within each other, thus we cannot cast this problem as binary classification.

where do we put this ? (Crü principally, D. Cristea, in The Right Frontier Constraint How Unconditionally. Proceedings of the Multidisciplinary Approaches to Discourse 2005 (MAD’05), and what do we say about it?
aside the problem of segmenting the texts into elementary discourse units, as the first stages of the annotation showed it was not difficult, and annotators could easily agree on the few discrepancies there were between segmentations. We had three annotators, each annotated 2/3 of the corpus and was paired with another annotator on a 1/3 of the corpus. They used segmentations they agreed upon before annotating rhetorical relations.

One of our three annotators is much less in agreement with the other two than these between themselves, and was found to be less reliable, so we present the best correlated pair of annotators. We estimated the common proportion of attachments of one with respect to the other as if the second one was the reference, which yields a F-score of 66%, for 279 common attachments. This is assuming attaching is a yes/no decision on every DU pair. But it should be noted this is not the way annotation works, as annotators try to cover minimally the text structure, and that some of these could be described in different syntactic ways, essentially with the use of complex units. The brutal estimation we give is thus likely to be an underestimation, and this raises the important issue of matching/comparing rhetorical structures. Refining this comparison is a work in progress, and should involve some kind of reasoning over the structures.

To give an indication of the problem, for a sequence of three EDUs (a),(b) and (c), where (a) is jointly elaborated by (b) and (c) with some coordination; one annotator could write the relations \{elaboration(a,b), elaboration(a,c), continuation(b,c)\} while the other chose to express the structure with a complex unit \[b-c\], and annotate only \{elaboration(a,\[b-c\]), continuation(b,c)\}. In semantic terms we could see these as equivalent (consider for instance that any part of an elaboration describes some sub-events of (a)) but in terms of agreement, there is only one common relation out of 2 or 3. See (Asher et al., 2011) for a preliminary study of what structural properties are needed to handle this issue.

Assuming the commonly attached pairs, the agreement on labels was then computed and yielded a Cohen kappa of 0.4 for the full set of 17 relations, which is a moderately satisfying agreement level. There is an important dispersion of annotations, and the majority class (entity elaboration) represents about 30% of the whole.

We also evaluated agreement on groups of relations, for instance the groups of coordinating versus subordinating relations, similar to the distinction between satisfaction-precedence and hierarchical relations in (Grosz and Sidner, 1986), for which we got a kappa of .57. Again, this raises the issue of equivalent rhetorical structures which could be ascribed to the same portions of text, and we are working on defining a satisfactory discourse graph matching.

What is involved here is a modelling of semantic consequences of rhetorical relations, and how they overlap for some relations (eg a “result” and a “narration” both entail a temporal ordering of the events they relate), which might explain some of the confusion between annotators and should be accounted for in the agreement measure. (Roze, 2011) has started to investigate this interplay of semantic consequences and what can be inferred from a combination of rhetorical relations.

### Table 4: A quantitative overview of annotated Multi-level Structures (a)

<table>
<thead>
<tr>
<th>corpus</th>
<th>ES</th>
<th>item</th>
<th>trigger</th>
<th>closure</th>
<th>enumeraTheme</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIK2</td>
<td>332</td>
<td>1639</td>
<td>296</td>
<td>34</td>
<td>167</td>
<td>232</td>
</tr>
<tr>
<td>LING</td>
<td>263</td>
<td>838</td>
<td>224</td>
<td>46</td>
<td>151</td>
<td>68</td>
</tr>
<tr>
<td>GEOP</td>
<td>234</td>
<td>716</td>
<td>180</td>
<td>43</td>
<td>120</td>
<td>187</td>
</tr>
<tr>
<td>ANNODIS</td>
<td>829</td>
<td>3193</td>
<td>700</td>
<td>123</td>
<td>438</td>
<td>487</td>
</tr>
</tbody>
</table>

### Table 6: Confusions wrt to the most common relation (elaboration)

<table>
<thead>
<tr>
<th>relation</th>
<th>vs elab</th>
<th>relation</th>
<th>vs elab</th>
</tr>
</thead>
<tbody>
<tr>
<td>attribution</td>
<td>-</td>
<td>explanation</td>
<td>5</td>
</tr>
<tr>
<td>background</td>
<td>6</td>
<td>flashback</td>
<td>1</td>
</tr>
<tr>
<td>comment</td>
<td>-</td>
<td>frame</td>
<td>6</td>
</tr>
<tr>
<td>conditional</td>
<td>-</td>
<td>goal</td>
<td>2</td>
</tr>
<tr>
<td>continuation</td>
<td>2</td>
<td>narration</td>
<td>-</td>
</tr>
<tr>
<td>contrast</td>
<td>4</td>
<td>parallel</td>
<td>1</td>
</tr>
<tr>
<td>e-elab</td>
<td>28</td>
<td>result</td>
<td>2</td>
</tr>
<tr>
<td>elaboration</td>
<td>25</td>
<td>temploc</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 7: Agreement on main sub-classes of relations

<table>
<thead>
<tr>
<th></th>
<th>None</th>
<th>Coordinating</th>
<th>Subordinating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinating</td>
<td>2</td>
<td>36</td>
<td>20</td>
</tr>
<tr>
<td>Subordinating</td>
<td>17</td>
<td>206</td>
<td></td>
</tr>
</tbody>
</table>

### 3.2. Multi-Level Structures

3.2.1. Two frequent and well-identified textual strategies

The annotation campaign clearly establishes the two structures as intuitive and quite easy to annotate, as was suggested early on by the satisfactory F-measures (2.3.). They are also seen to be frequent, and to occur at different levels of text structure, indicating that they are relevant patterns for studying the complexity of discourse organisation. They are found in all three sub-corpora in the ANNODIS corpus: 5 to 12 topical chains per 10000 words, and 11 to 18 enumerative structures depending on the sub-corpus. Topical chains occupy on average 15% of the texts’ surface, against 43% for enumerative structures. Enumerative structures appear at different levels of granularity, with every level of the text’s structure potentially concerned: they can stretch over
several sections, several paragraphs within a section, or be bounded within a single paragraph. 4 On the basis of these initial observations, both structures appear as basic strategies to which writers resort frequently in different genres of expository/argumentative texts. The following subsections focus on further results concerning enumerative structures (ESs).

### 3.2.2. A formal typology of enumerative structures

The following typology is the one that best explains the variations in composition and cue usage in enumerative structures: the different types are described in terms of their interaction with document structure at the different granularity levels that we have just mentioned. In ESs of Type 1, labelled ‘Headed sections’, each item corresponds to a section (or subsection). Type 2 ESs are formatted lists. They are defined solely in terms of specific typographical and layout features (bullet points or numbers). They may be local formatted lists composed of only two items or large-scale lists of up to 48 items covering an entire section. Type 3 ESs are multiparagraphic structures. On the most local level, type 4 depicts ESs that are inserted inside a paragraph or corresponding exactly to a paragraph.

Concerning the main characteristics of these four visual types of ESs, some simple statistical measures provide the following interesting significant correlations: Types 1 and 2 are characterised by a higher cardinality (3.8 items on average against 3) and a higher presence of triggers; enumeraThemes are more often present in Type 2 ESs and less often in Type 1 ESs; closures are significantly less frequent in Type 1 ESs. Cross-corpus comparisons are shown on table 8. These figures show that significant differences appear between corpora. Wikipedia articles are characterized by a larger amount of type 1 and particularly type 2 ESs, whereas local ESs are particularly present in the other two corpora, which resort less to multisection ESs.

Concerning the main characteristics of these four visual types of ESs, some simple statistical measures provide the following interesting significant correlations: Types 1 and 2 are characterised by a higher cardinality (3.8 items on average against 3) and a higher presence of triggers; enumeraThemes are more often present in Type 2 ESs and less often in Type 1 ESs; closures are significantly less frequent in Type 1 ESs. Cross-corpus comparisons are shown on table 8. These figures show that significant differences appear between corpora. Wikipedia articles are characterized by a larger amount of type 1 and particularly type 2 ESs, whereas local ESs are particularly present in the other two corpora, which resort less to multisection ESs.

Concerning the main characteristics of these four visual types of ESs, some simple statistical measures provide the following interesting significant correlations: Types 1 and 2 are characterised by a higher cardinality (3.8 items on average against 3) and a higher presence of triggers; enumeraThemes are more often present in Type 2 ESs and less often in Type 1 ESs; closures are significantly less frequent in Type 1 ESs. Cross-corpus comparisons are shown on table 8. These figures show that significant differences appear between corpora. Wikipedia articles are characterized by a larger amount of type 1 and particularly type 2 ESs, whereas local ESs are particularly present in the other two corpora, which resort less to multisection ESs.

### 3.2.3. Towards a functional typology of enumerative structures

As stated in 2.3, coders were asked to annotate lexical expressions referring to the co-enumerability criterion, the enumeraTheme, or underlying ‘theme’ of the enumerative structure. A first typology of these annotations distinguishes three types: a concept (as in ‘the theory is based on three principles’), an entity (as in ‘individuals are split up into 3 groups’) or a textual object (as in ‘this paper consists of four sections’). The vast majority (80%) of ESs concern concepts, against 9% of entities and 7.5% of textual objects. Though this typology is only preliminary and the ‘concept’ class in particular needs refining, this initial result suggests that ESs are predominantly used to create new categories via discourse rather than to refer to pre-existing categories or as metadiscourse.

### 4. Intersecting the bottom-up and top-down approaches

Given the top-down approach’s hypothesis that high level structures affect the interpretation of other structures within their scope, we expect that top-down annotated structures will place constraints on the graph constructed via the bottom-up method. Extracts of a subset of the texts in the WIK2, LING and GEOP parts of the corpus were subject to both top-down and bottom up annotation methods, see table 9.

<table>
<thead>
<tr>
<th>sub-corpus</th>
<th>Nb texts</th>
<th>Nb excerpts</th>
<th>N words</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIK2</td>
<td>9</td>
<td>12</td>
<td>4908</td>
</tr>
<tr>
<td>LING</td>
<td>3</td>
<td>3</td>
<td>1116</td>
</tr>
<tr>
<td>GEOP</td>
<td>3</td>
<td>3</td>
<td>1340</td>
</tr>
</tbody>
</table>

Table 9: Part of the ANNODIS corpus at the intersection of the two approaches

While a full understanding of the constraints induced by high level structures remains something for future study, several hypotheses already seem promising. 1) the macro-level structures can serve to guide CDU construction. As CDUs do not overlap, we predict that there should be no CDU that does not properly cover CDUs isolated by macro-methods. 2) macro-level structures such as enumerations can determine the semantic value of certain discourse markers like *puis*. If the overall structure, for instance, enumerates arguments in support of some hypothesis, a use of *puis* in the enumeration of those arguments should only be taken as indicating an instance of one of the arguments in the list, not a temporal sequence (which is what *puis* is typically used to do in the bottom-up approach). We hope to study constraints like these and enlarge the coverage of the dually annotated corpus in future work.

### 5. Conclusion

The ANNODIS corpus incorporates two levels of discourse annotation: a bottom-up type annotation of elementary and complex discourse units along with the coherence relations that connect those structures, and a top-down annotation of high level discourse structures such as enumerative structures. The bottom-up annotations of the ANNODIS corpus differ from those in other annotation efforts that give a discourse structure for an entire text. For example, compared to the RST corpus, ANNODIS incorporates a wider array of structures; it also distinguishes between complex discourse units and EDUs explicitly, which RST arguably does.
not, unless one adopts Marcu’s nuclearity principle (Marcu, 2000). We plan to investigate how the nuclearity principle relates to ANNODIS structures in future work. Discourse pop-ups for non-contiguous spans of text are also explicitly marked in the ANNODIS corpus. In relation to PDTB, the ANNODIS corpus creates full discourse structures instead of providing simply coherence relations between contiguous phrases. Finally, this corpus has led to the creation of various discourse-oriented tools (e.g., a segmenter) and has served to validate the right frontier constraint of discourse. We are currently working on making more explicit the differences of this framework in relation to other more well known frameworks or corpora. The creation of a discourse parser is among our immediate goals as well.

6. References


R. Power, D. Scott, and N. Bouayad-Agha. 2003. Docu-


