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

Stergos Afantenos[1], Nicholas Asher[1], Farah Benamara[1], Myriam Bras[2], Cécile Fabre[2], Mai Ho-dac[2], Anne Le Draoulec[2], Philippe Muller[1], Marie-Paule Péry-Woodley[2], Laurent Prévot[3], Josette Rebeyrolle[2], Ludovic Tanguy[2], Marianne Vergez-Couret[2], Laure Vieu[1]

[1]IRIT, Univ. Toulouse, France; [2]CLLE, Univ. Toulouse, France; [3]LPL, Aix-Marseille Univ., France
(authors are in alphabetical order)

Abstract

This paper describes the ANNODIS resource, a discourse-level annotated corpus for French. The corpus combines two perspectives on discourse: a bottom-up approach and a top-down approach. The bottom-up view incrementally builds 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 followed here involves 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 also serves as a source of empirical evidence for discourse theories. We present here two first analyses taking advantage of this new annotated corpus, one that tested hypotheses on constraints governing discourse structure, and another that studied the variations in composition and signalling of multi-level discourse structures.

Keywords: Corpus linguistics, discourse structure, document structure, corpus annotation

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, as far as we know, the first resource of this kind. 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 realised 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 novel synergies. The top-down approach provides a macro level organisation 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 has 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 thousand words each).

In order to provide a diversified corpus, we selected texts showing 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 *Est Républicain*, encyclopedia articles (from the French Wikipedia), linguistics research papers (from *Congrès Mondial de Linguistique Française*) and international relation reports (from the *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 documents

(sections, headings, paragraphs, etc.) and is presented on a three level scale; this parameter is determined by the source.

Table 1, page 3, summarises the composition of the corpus, along with the number and total size of texts for each category. The first two rows describe the bottom-up part of the corpus, the last three the top-down part. However, there is some overlap between these two subsets, as some of the top-down texts have been annotated according to both methods, as presented in section 4.

Every text is protected by a Creative Commons license that allows us to make the Annodis corpus freely available for research purposes; this aspect played an important role in the selection of the sources.

Although the two annotation campaigns were based on different approaches to discourse organisation (respectively bottom-up and top-down), and required different kinds of text (see section 4.), they proceeded in similar ways: with the help of an annotation manual, three naive coders annotated objects in texts using the Glozz interface (Mathet and Widlöcher, 2009)[1]. Glozz is an annotation tool originally created for the annotation of the ANNODIS resource, which implements a generic model allowing the annotation of units, relations and schemata. It provides advanced text-visualisation facilities, whereby texts can be displayed as real-life documents (with visual signalling such as paragraph breaks, several levels of headings, bullets/numbered lists, etc.) and pre-marked features highlighted in order to assist annotation. The next two subsections describe the two annotation approaches and give an overview of the data annotated for each.

2.2. Bottom-Up Approach: Rhetorical Relations

The bottom-up approach used both naive and expert annotators in three distinct phases of annotation. During the first, preliminary phase, two graduate-level students annotated 50 documents. We used their input in order to create the annotation manual used in 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.

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. 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 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 contraints on attachment (the so-called "right frontier" of discourse structure, see section 3.1.2.). 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 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.

[Principes de la sélection naturelle.]_1 [La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,]_2 repose sur trois principes:]_3 [1. le principe de variation]_4 [2. le principe d'adaptation]_5 [3. le principe d'hérédité]_6

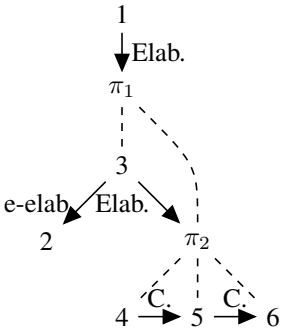

Figure 1: **An example of discourse graph**. The nodes correspond to discourse units; the EDUs are represented by their numbering; the CDUs start with π. Dotted edges represent inclusion to a CDU while edges with arrows represent rhetorical relations. Elab. = Elaboration, e-elab = Entity Elaboration,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

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 takes elements largely common to the theories on the market—Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), the Linguistic Discourse Model (LDM) (Polanyi et al., 2004) the GraphBank model (Wolf and 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). 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 recursively 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, nuclei being more important for

[1] http://www.glozz.org/

Id	Source	Genre	Type	Document structure	Texts	Tokens
NEWS	*Est Républicain*	news	narrative	low	39	10K
WIK1	*Wikipedia* excerpts	encyclopedia	expository	low	30	11K
WIK2	*Wikipedia*	encyclopedia	expository	high	30	231K
LING	*CMLF-08*	research	expository	medium	25	169K
GEOP	*IFRI (international relations)*	reports	argumentative	medium	32	266K
Total					156	687K

Table 1: Breakdown of the Annodis corpus

the relation.[2] 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 that is more expressive than that of other theories (Danlos, 2007) and allows for long distance attachments, pop-ups and even some crossed dependencies. 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 (Hovy and Maier, 1993). 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 the 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, ATTRIBUTION, BACKGROUND, FLASHBACK, FRAME, TEMPORAL-LOCATION, ELABORATION, ENTITY-ELABORATION, COMMENT. Most of these are self-explanatory (cf also (Asher and Lascarides, 2003; Vieu et al., 2005; Prévot et al., 2009)).

Table 2 shows the number of discourse units annotated in the corpus, with 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. Section 3.1.3. presents more information on the inter-annotator agreement.

[2]Some relations can be multi-nuclear, meaning that all of their arguments (spans) are important for that relation.

	corpus total	NEWS	WIK1
Nb Texts	87	39	42
Nb words	28146	9768	17330
EDU	3188	1159	1949
CDU	1395	510	829

Table 2: Discourse unit counts in expert annotations

2.3. Top-down Approach: Multi-level Structures

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). To translate this view into a realistic annotation program, 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 whereby text elements are presented as having equal status with regard to a specific interpretation criterion. A variety of devices, which are often combined, imply a similarity between the items of an enumeration: formatting, typography, numbering, syntactic parallelism, lexical item introducers. In an enumerative structure, two optional segments may prefix and conclude the list of enumerated items: a trigger and a closure. Enumerative structures thus have an internal organisation consisting of 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*. As inherently signalled textual motifs, enu-

	total (Nb)	(%)	NEWS (%)	WIK1 (%)		total (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
Eelab	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	3355	100	1203	2034					

Table 3: Discourse relations of the expert annotations

Le LAF, rédigé en collaboration avec Igor Mel'cuk, est un travail 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 destiné, en tout premier lieu, aux enseignants de langue et un 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 lexicographiques), **le LAF** peut être rapproché de Picoche (2007). Il est intéressant de constater que le travail d'interfaçage 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 lexicales (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 dériver de la base source DiCo d'autres « produits », comme celui dont il va maintenant être question.	TC

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

II) Des orientations d'action Les orientations proposées peuvent être regroupées autour de quatre thèmes.	ES	TRIGGER
- Mieux organiser notre politique étrangère dans la région ce qui passe, notamment, par la mise en place de structures permettant [...].		ITEM 1
- Accentuer notre coopération avec des partenaires d'influence, notamment en établissant une coopération renforcée avec certains [...].		ITEM 2
- Manifester notre souci de voir émerger des systèmes démocratiques dans la région en développant une politique d'influence auprès des [...].		ITEM 3
- Contribuer plus efficacement à la solution des principales crises régionales, ce qui comporterait les actions suivantes : [...].		ITEM 4
En conclusion, les turbulences qui affectent le moyen orient ont atteint un niveau de haute intensité qui représente, pour les pays occidentaux et, plus spécialement, pour l'Europe, de grands risques, notamment [...].		CLOSURE

Figure 3: ES – Enumerative Structure – covering a whole subsection: the heading together with the opening paragraph announce that the following text will list four "themes"; next, the identity of presentation of the four items signals their similarity with respect to this co-enumerability criterion; finally, the last paragraph of the subsection closes the enumeration.

merative 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 signalling; finally, their ability to occur at vastly different levels of text granularity is of particular interest in exploring the articulation between levels of text organisation.

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) was automatically carried out on the POS-tagged and syntactically analysed texts, with TreeTagger and SYNTEX (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), punctuation (e.g. paragraphs ending with [:], punctuational motifs such as [: ...; ...; and/or ...]), as well as lexico-syntactic features. Via specifically-built lexica and local grammars, the following lexico-syntactic features were pre-marked: coreferential and topical expressions (Cornish, 1999) e.g. pronouns and lexical reiterations; item introducers (Hempel and Degand, 2008) e.g. *firstly*, *finally*, *the first X*, *on the other hand,* ; predictive 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.). It must also be pointed out that a specially designed style-sheet enabled the annotation to be performed on naturalistic text, i.e. with preserved layout.

The human annotation then proceeded as follows: using the Glozz interface, coders detected ESs and TCs by scanning the text with the help of visual layout and highlighted pre-marked features. For each structure detected, they indi-

cated the boundaries of its segments and sub-segments, and, in the case of ESs, the enumeraTheme, i.e. the expression specifying the co-enumerability criterion. Finally, they annotated the cues 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).

We organized the annotation program 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. Pairs of annotations were compared as regards segments of text concerned, sub-segments for ESs and main referent for CTs. Agreement was calculated in terms of F-measure, with results of 0.7 for ESs and 0.65 for TCs . 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 texts[3] (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 for the three sub-corpora:

corpus	added features	validated premarked features
WIK2	1677	2428
LING	937	708
GEOP	1130	993
Total	3744	4129

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

3. First Experiments and Analyses

3.1. Rhetorical Relations

A corpus of texts annotated with discourse structure allows for a number of empirical studies on semantic and pragmatic phenomena. It also feeds work on automated prediction of discourse structures. We present here the efforts that are under way within the project, which have already yielded 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.[4] 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 $B = \{\texttt{left}, \texttt{right}, \texttt{both}, \texttt{nothing}\}$. 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. The features that we used were mostly based on surface and morpho-syntactic information.

For our classifier, we used a regularized maximum entropy (MaxEnt) model (Berger et al., 1996). The classification was followed by a post-processing enforcing well-balanced segments. Ppost-processing yielded an F-measure of 0.733 for the EDUs as a whole. For more details, see (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) 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 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, 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 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.

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

[4] In contrast to other theories EDUs in SDRT can be embedded within each other, thus we cannot analyze this problem using a binary classification.

corpus	ES	item	trigger	closure	enumeraTheme	TC
WIK2	332	1639	296	34	167	232
LING	263	838	224	46	151	68
GEOP	234	716	180	43	120	187
Total	829	3193	700	123	438	487

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

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, and that all decisions are independent. But it should be noted this is not true in practice, as annotators can express similar structures in different 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, consider a sequence of three EDUs (a), (b) and (c), where (a) is jointly elaborated by (b) and (c) with some coordination between (b) and (c); 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 subevents 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.

Considering commonly attached pairs only, 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. As seen table 3, there is an important dispersion of annotations.

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.

	None	Coordinating	Subordinating
Coordinating	2	36	20
Subordinating		17	206

Table 6: Agreement on main sub-classes of relations

3.2. Multi-Level Structures

The first results of our—mostly quantitative—analysis of the data resulting from the annotation campaign are of interest in terms of the following issues: the relations between types of signalling devices ("visual" vs lexico-syntactic), the articulation between modes of textual organisation, and the functions of the annotated structures in discourse.

3.2.1. Two frequent and well-identified textual strategies

Our annotation method assumed a functional equivalence between lexico-syntactic signalling devices and "visual" devices (positional, typographical), which linguists have generally ignored: they were presented together in the annotation manual as well as in the interface. The annotation campaign clearly establishes that these devices make ESs and topical chains intuitive and easy to annotate, as confirmed by the satisfactory F-measures (section 2.3.). They are also frequent and occur at different levels of text structure, indicating that they are relevant patterns for studying the complexity of discourse organisation. We found them in all three sub-corpora used for the second task: 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.[5] 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).

[5]The annotation scheme for topical chains restricted the annotation to segments covering no more than one section (Fig 2 shows a one section topical chains), which means that potential very high-level chains were not annotated.

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 7. These figures show that significant differences appear between corpora. Wikipedia articles are characterised 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.

Corpus	Types of ESs			
	Headed sections	Formatted lists	Multi-§ ES	intra-§ ES
WIK1	19.3	39.1	20.8	20.8
LING	9.1	23.2	26.6	41.1
GEOP	6.8	10.3	20.9	62

Table 7: **Distribution of ESs types (percent)**

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: concept (as in "the theory is based on three principles"), concrete entity (as in "individuals are split up into 3 groups") or textual object (as in "this paper contains four sections"). The vast majority (80%) of ESs were analyzed as having "concepts" as co-enumerability criteria, against 9% of "concrete 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 8.

sub-corpus	Nb texts	Nb excerpts	N words
WIK2	9	12	4908
LING	3	3	1116
GEOP	3	3	1340

Table 8: 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 doubly 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 structurefor 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 empirically validate the right frontier constraint

of discourse. We are currently working on making more explicit the differences between this framework and other more well known frameworks or corpora. The creation of a discourse parser is among our immediate goals as well.

As for the annotated multi-level structures, they constitute an original resource for studying text organisation strategies and signalling, especially at higher levels of textual granularity. The availability of a diversified corpus enriched with exhaustive human annotations of these structures and their cues opens up the possibility of using data-mining techniques to examine how cues interact, and how cue combinations vary according to genre and text-type. Exploring the hypothesis that lexical markers are only the most visible part of complex discourse markers combining lexical, syntactic, positional and typographical features, we are currently working on the identification of "cuesets" for the different multi-level structures, with potential applications in automatic text synthesis and document navigation.

6. References

S. Afantenos and N. Asher. 2010. Testing SDRT's Right Frontier. In *Proceedings of COLING 2010*, pages 1–9.

S. Afantenos, P. Denis, P. Muller, and L. Danlos. 2010. Learning Recursive Segments for Discourse Parsing. In *Proceedings of LREC 2010*.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

N. Asher, A. Venant, P. Muller, and S. D. Afantenos. 2011. Complex discourse units and their semantics. In *Constraints in Discourse (CID 2011)*, Agay-Roches Rouges, France.

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

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71, March.

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

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

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

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

L. Danlos. 2007. Strong generative capacity of RST, SDRT and discourse dependency DAGs. In A. Benz and P. Kühnlein, editors, *Constraints in Discourse*. Benjamins.

K. Forbes, E. Miltsakaki, R. Prasad, A. Sarkar, A. K. Joshi, and B. L. Webber. 2003. D-ltag system: Discourse parsing with a lexicalized tree-adjoining grammar. *Journal of Logic, Language and Information*, 12(3):261–279.

G. Francis. 1994. Labelling discourse: an aspect of nominal-group lexical cohesion. In M. Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. Routledge: London.

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

B. Grosz and C. L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-September.

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

S. Hempel and L. Degand. 2008. Sequencers in different text genres: Academic writing, journalese and fiction. *Journal of Pragmatics*, 40:676–693.

E. Hovy and E. Maier. 1993. Parsimonious or profligate: How many and which discourse structure relations? Unpublished manuscript, available at `http://www.isi.edu/natural-language/people/hovy/papers/93discproc.pdf`.

W. Mann and S. Thompson. 1987. Rhetorical structure theory : a theory of text organization. Technical report, Information Science Institute.

D. Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

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

L. Polanyi, C. Culy, M. van den Berg, G. L. Thione, and D. Ahn. 2004. A rule based approach to discourse parsing. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 108–117, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

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

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

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proceedings of LREC'08*, Marrakech, Morocco, may. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/.

L. Prévot, L. Vieu, and N. Asher. 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.

C. Roze. 2011. Towards a Discourse Relation Algebra for Comparing Discourse Structures. In *Constraints in Discourse (CID 2011)*, Agay-Roches Rouges, France.

L. Vieu, M. Bras, N. Asher, and M. Aurnague. 2005. Locating adverbials in discourse. *Journal of French Language Studies*, 15(2):173–193.

F. Wolf and E. Gibson. 2005. Representing discourse coherence: A corpus based study. *Computational Linguistics*, 31(2):249–287.