# A corpus-driven approach to discourse organisation: from cues to complex markers

Marie-Paule Péry-Woodley, Lydia-Mai Ho-Dac, Josette Rebeyrolle, Ludovic Tanguy, Cécile Fabre

Dialogue & Discourse 8(1) (2017) 66–105 doi: 10.5087/dad.2017.103

# A corpus-driven approach to discourse organisation: from cues to complex markers

**Marie-Paule Péry-Woodley** PERY@UNIV-TLSE2.FR
*CLLE, Université de Toulouse*
*CNRS, UT2J, France*

**Lydia-Mai Ho-Dac** HODAC@UNIV-TLSE2.FR
*CLLE, Université de Toulouse*
*CNRS, UT2J, France*

**Josette Rebeyrolle** REBEYROL@UNIV-TLSE2.FR
*CLLE, Université de Toulouse*
*CNRS, UT2J, France*

**Ludovic Tanguy** TANGUY@UNIV-TLSE2.FR
*CLLE, Université de Toulouse*
*CNRS, UT2J, France*

**Cécile Fabre** CECILE.FABRE@UNIV-TLSE2.FR
*CLLE, Université de Toulouse*
*CNRS, UT2J, France*

**Editor:** Maite Taboada

Submitted 03/2016; Accepted 12/2016; Published online 01/2017

## Abstract

This paper reports on an experiment implementing a data-intensive approach to discourse organisation. Its focus is on enumerative structures envisaged as a type of textual pattern in a sequentiality-oriented approach to discourse. On the basis of a large-scale annotation exercise calling upon automatic feature mark-up alongside manual annotation, we explore a method to identify complex discourse markers seen as configurations of cues. The presentation of the background to what is termed "multi-level annotation" is organised around four issues: linearity, complexity of discourse markers, top-down processing, granularity and the multi-level nature of discourse structures. In this context, enumerative structures seem to deserve scrutiny for a number of reasons: they are frequent structures appearing at different granularity levels, they are signalled by a variety of devices appearing to work together in complex ways, and they combine a textual role (discourse organisation) with an ideational role (categorisation). We describe the annotation procedure and experimental framework which resulted in nearly 1,000 enumerative structures being annotated in a diversified corpus of over 600,000 words. The results of two approaches to the rich data produced are then presented: firstly, a descriptive survey highlights considerable variation in length and composition, while showing enumerative structure to be a basic strategy resorted to in all three sub-corpora, and leads to a granularity-based typology of the annotated structures; secondly, recurrent cue configurations—-our "complex markers"—-are identified by the application of data mining methods. The paper ends with perspectives for further exploitation of the data, in particular with respect to the semantic characterisation of enumerative structures.

**Keywords:** discourse structures, discourse markers, corpus linguistics, corpus annotation, data mining

## 1. Introduction

Texts can be seen as the result of squeezing complex hierarchical structures into a largely linear format. Understanding a text entails constructing a representation of the underlying structures. A major challenge in the study of written discourse is to identify the signals which guide readers in the process of constructing this representation. Depending on one's theoretical underpinning and focus, signals may be seen as discourse construction devices, as metadiscourse, as reading or processing instructions, as traces of the writers cognitive processes, or as cues revealing the authors intentions. The study presented in this paper sets up a data-intensive methodology whereby signals "emerge" from the systematic analysis of a large set of annotated structures. Its aim is the empirical characterisation of configurations of cues signalling a particular discourse pattern: enumerative structures. As these structures can concern text spans of any size, the perspective is described as multi-level. The study relies on the systematic annotation of structures in a corpus of French language texts, and on the application of data mining methods to detect emergent complex discourse markers. While in terms of methodology it belongs in corpus linguistics and natural language processing, its theoretical foundations are to be found in functional linguistics, in psycholinguistics and in research on the visual dimension of texts. We chose to start from what may be seen as the most basic among the notions called upon to account for text/discourse organisation: linearisation, continuity vs. discontinuity (the fundamental question behind discourse segmentation), and discourse patterns.

The arguments for this "back to basics" approach are given in the next section, organised around four issues: the linearity constraint, the non-discrete nature of discourse markers, the importance of top-down processing, granularity and the multi-level nature of discourse structures. These constitute the foundation for the choice of enumerative structures for annotation, the rationale for which is given in Section 3, followed in Section 4 by the annotation model and method, from corpus preparation procedures to the manual annotation of structures and cues. In Section 5, a descriptive survey of nearly 1,000 annotated structures leads to a proposal for a granularity-based typology, and to an analysis of genre-related variations. Finally, Section 6 presents the recurrent cue configurations made apparent by the application of data mining techniques to the rich annotated data.

## 2. Multi-level annotation in the ANNODIS project: preliminaries

The research presented here started with the ANNODIS annotation project, which can be described as a large-scale discourse-level annotation experiment calling upon different discourse models and different genres of written French language texts[1]. The project comprised two distinct approaches, respectively labelled bottom-up and multi-level. Bottom-up and multi-level annotations were applied to different corpora, for reasons which are explained in 4.2.2, but a set of texts was annotated in both frameworks to allow a direct comparison of the approaches (ANNODIS_duo, see 5.3). The bottom-up annotation, conducted according to Segmented Discourse Representation Theory (Asher, 1993), focused on the identification of rhetorical relations. The multi-level annotation—-the main focus of this paper—-took a less well-chartered path, which the present section aims to describe and justify in the form of basic propositions underlying the choice of objects to annotate, the annotation method, as well as the questions asked of the annotated corpus.

1. Project funded by the Humanities and Social Sciences Programme of the French National Research Agency (ANR Appel Corpus 2007) (see Péry-Woodley et al., 2011; Afantenos et al., 2012, for complete descriptions). The ANNODIS resource is available under a creative commons licence `http://redac.univ-tlse2.fr/corpus/annodis`.

### 2.1 Linearisation is a problem

Language is linear, while mental representations are not (or not necessarily). This, as many authors have pointed out (Levelt, 1981; Gernsbacher, 1995, 1997; Heurley, 1997, *inter alia*), can be seen as problematic insofar as "in text, a multidimensional discourse model is squeezed into a linear form. Linearity requires the writer to produce each textual unit in turn, and processing constraints demand short units of meaning. Yet the mental representation on which the discourse is based is not a succession of facts or ideas which can each be expressed in one sentence. This is where discourse organisation comes in..." (Ho-Dac and Péry-Woodley (2009), echoing Levelt (1981)). Few approaches to expository discourse, however, focus on linearisation and its inescapable consequence—-sequentiality, i.e. the segmentation of discourse into subsequent text spans.

Goutsos (1996, 1997) is one author who argues for a theory of sequential relations, in which he sees "an autonomous source of text connectivity" (ibid. 502). He describes most approaches as favouring a what-perspective—-"what is taking place in discourse" ("propositional or semantic content—-over a how-perspective—-"which would focus on the structuring rather than the individual units of text" (Goutsos, 1996, p. 503). The macrostructure or story grammar approach to discourse coherence (van Dijk, 1980) is an example of the what-perspective, as are, largely, models relying on the notion of rhetorical relations (Rhetorical Structure Theory (Mann and Thompson, 1988), Segmented Discourse Representation Theory (Asher, 1993), *inter alia*). Goutsos' how-perspective has its roots in functional linguistics, in the notion of "information packaging" (Chafe, 1976, 1994), in Halliday's textual metafunction (Halliday, 1977/1983), in the notion of textual strategy (Enkvist, 1985; Virtanen, 1992). A number of researchers in the field of automatic text generation, especially in the RST "sphere", have developed models based on a distinction related to the one Goutsos proposes: in particular Virbel, via his Text Architecture Model based on the notion of "textual object" (Virbel, 1989, 2015; Lemarié et al., 2008), and Power et al., who argue that what they call "abstract document structure" is a separate descriptive level in the analysis and generation of written texts (Power et al., 2003). A distinction does exist within RST between subject-matter and presentational relations, a distinction which Taboada and Mann (2006) associate with Goutsos' proposals [p. 443]. But Power et al. go further by questioning the ambiguity in RST between text spans and the meanings of these spans in the attribution of relations, and call for a clear distinction between document structure and rhetorical structure, which they claim are as distinct as syntax from semantics (Power et al., 2003, p. 245).

These authors have in common a central concern with text segmentation, and how it is signalled (i.e. with what signalling devices). They do not primarily focus on the nature of relations between text segments—-the central concern for models of discourse organisation based on discourse relations. In Goutsos' model, the two fundamental relations between text spans are simply continuity and discontinuity (or shift), text being seen as a "periodic alternation of transition and continuation spans" (Goutsos, 1996, p. 501). Continuity applies by default, and therefore can be implicit, whilst discontinuity requires some form of signalling, i.e. the presence of linguistic devices that "function as cues to the reader", "help[ing] the reader assign the utterance in which they occur to a continuation or a transition space" (Goutsos, 1996, p. 517). The next section sketches out the conception of discourse organisation signals which is embodied in the present study.

### 2.2 Signalling discourse organisation is "a struggle between different forces"

At any given time, linguistic choices in text production are influenced by several principles concurrently at play. These choices are, in Enkvist's words, "the outcome of a conspiracy or a struggle between the different forces that affect the linearization of discourse" (Enkvist, 1985, p. 321). We shall use an example to explain why Enkvist's image seems relevant:

**Example 1** *from a policy oriented text published by IFRI (French Institute for International Relations)*[2]

> ***New budgetary cuts*** *[section heading]*
> *Let us now look at the effect of the crisis as things stand at present. [...]*
> *In the United Kingdom , the defence budget, which amounted to 44.5 billion in addition to spending relating to external operations in Afghanistan and Iraq, is to be cut by [...]*
> *In Germany , debate is raging over whether or not to abolish national service, which would reduce troop numbers from 250,000 to [...]*
> *In Austria , a question mark is hanging over the military service and most of the country's tanks have been withdrawn from service. [...]*
> *In Greece, the defence budget will be amputated by [...]*

On reading Example 1, one is immediately aware of the presence of a number of paragraph-initial adverbials (*In the United Kingdom*, *In Germany*, etc.). Time and space adverbials are amongst potential sequentiality cues which have also been considered good segmentation markers by researchers working in a what-perspective: they can be associated with topic shifts (Piérard and Bestgen, 2006), and they introduce a new interpretation criterion projecting forward (Charolles et al., 2005). Adopting a how-perspective, we would argue that what is significant about these adverbials is that there are four of them in relatively short succession, exhibiting strong parallelism: paragraph-initial prepositional phrases (*In* + name of country) followed by a comma. Together, they form an identifiable pattern, and recognising this pattern is in our view very much part of understanding what the text is about. Now, a series of paragraph-initial sequencers (*Firstly, Secondly,...*) instead of adverbials, or adverbials of time instead of space, would create a functionally similar pattern. A sequence of four non-initial, non-detached adverbials, on the other hand, would definitely not realise the same text strategy (Virtanen, 2004), and would not be perceived in the same way. In this perspective, features that are often overlooked in discourse organisation research must be fully taken into account, in particular layout and punctuation: the paragraph breaks, as well as the commas following each prepositional phrase, are clearly determining features. The pattern at once delineates the items as discontinuous and brings them together—-because of their parallelism—-into a higher level span (see Figure 2 below).

Viewed from a what-perspective, the four adverbials in Example 1 introduce spatial criteria which are essential for the interpretation of subsequent text: everything said after adverbial $n$ and before adverbial $n + 1$ only applies in (is only true for) the spatial area designated by the adverbial. We find unconvincing the dichotomy sometimes found in the literature on textual metadiscourse between propositional and non-propositional textual material (or primary and secondary discourse) (see Ho-Dac et al., 2012). The adverbials in Example 1 are both at once: they have both an ideational and a textual role, and they are noteworthy for both the what- and the how- perspectives. This duality takes us back to Enkvist's remark about "a struggle between different forces": given that at any one time several processes are concurrently going on in text (producing content, organising text), signalling devices must be expected to be largely multifunctional, since they may be shared by several processes.

2. `http://www.ifri.org/downloads/europevisions7ojehin.pdf`

The perspective sketched out here challenges the view of discourse organisation as signalled in text primarily via specialised (lexical) discourse markers. Along the lines of authors such as Marcu (Marcu, 2000, 2006), we would argue that discourse markers are likely to be more eclectic and less discrete, i.e. to come in the form of bundles of cues in which most of the time no single element is either necessary or sufficient. If, as suggested in the analysis of Example 1, and in line with a number of authors (Virtanen, 1992, 2004; Hasselgård, 2010), adverbials only function as sequentiality markers when associated with other features (e.g. positional or punctuational features) and/or occurring in a series (Ho-Dac and Péry-Woodley, 2009), the search for discourse markers is redefined as a search for recurrent cue configurations. A similar approach is adopted in recent research on the signalling of implicit discourse relations (Taboada and Das, 2013). In their study, Taboada and Das stress that "we need to move beyond the signalling by discourse markers (...) in order to understand how relations are processed, and in order to extract them automatically" (idem, p.250). The authors propose annotating a wide range of cues in order to discover combined and multiple signals of discourse relations (idem, p.250), an objective close to our own search for complex discourse markers.

### 2.3 Top-down processing influences discourse interpretation

Discourse markers are usually seen as bearers of instructions. For instance, connectives carry the instruction to link two text segments (or the propositional meanings they convey) via a particular relation. This conception is rooted in a bottom-up view of discourse interpretation as a unit by unit construction. We focus here on movement in the opposite direction, proposing that in text processing—-particularly in the case of expository text—-there is also an immediate perception of high-level signals, a Gestalt-like grasp of large-scale textual patterns, which in turn influences the step by step interpretation process (Asher et al., to appear). This interest in top-down processing relates to research in neighbouring fields: cognitive psychologists and psycholinguists have studied the effect of headings and sub-headings, and of layout features such as paragraph breaks, on reading comprehension and recall (Lorch and Lorch, 1996; Lemarié et al., 2012, 2008; Heurley, 1997); computational approaches to document generation and understanding, in their concern with linearisation (cf. Section 2.1), have sought to define principles governing the physical presentation of text (Power et al., 2003; Bateman et al., 2001; Virbel, 1989, 2015). We use the term "textual pattern" to suggest that this top-down processing may have more to do with pattern recognition than with a compositional meaning-construction process. Signalling is part and parcel of the definition of textual patterns, since they are characterised by their ability to be readily perceived by readers. There can be no such thing as an implicit textual pattern.

### 2.4 Discourse structures are multi-level

In our attempt to draw attention to top-down processing, we referred to *high-level* signals and *large-scale* textual patterns. More precisely, the signals and textual patterns in question are typically multi-level and apply recursively, and these are properties which are of great interest to us. The textual pattern which is the focus of this paper, enumerative structures, will be seen to range from a few lines to whole sections of text, and to allow several levels of embedding. This multi-level property is clearly related to how the text spans delimited by discourse structures interact with document structure segmentation (sections and sub-sections, paragraphs). Example 1 illustrates such an interaction between two modes of organisation, where place adverbials and paragraph segmentation

may be seen as signalling the items of an enumeration. The approach therefore implies attention to document structure (Power et al., 2003), and its signals. Visual signals of document structure are considered as fully-fledged discourse features with the potential to combine with other cues to form what we call complex discourse markers.

The previous section has allowed us to clarify our general objective in the light of the basic tenets of our approach. We can now turn to the method specifically set up to identify these complex discourse markers, which involves automatic feature-tagging (Section 4.1.1), manual annotation of textual patterns (Section 4.1.2) and data mining techniques to identify correlations (Section 6). The method is designed for long expository texts which differ noticeably from newspaper articles in terms of discourse organisation (cf. Section 4.2.2). As described in Section 6.3, this method has made it possible to identify complex discourse markers made up of cues appearing in series or in specific patterns. We also insist on the role of genre in determining what cues are used, or in shifting the balance of interpretation of particular sets of cues. The first stage in the description of the method is to present the context—-the ANNODIS multi-level annotation experiment.

## 3. An annotation experiment to implement a multi-level approach to discourse

In order to observe diverse structuring modes, including at high levels of organisation, we devised an annotation experiment to be carried out on lengthy non-narrative texts, organised into three distinct sub-corpora so as to allow potential genre-related variation to emerge. In accordance with the approach presented above, the annotation is not based on predefined markers: the identification of cue configurations functioning as discourse markers is an expected outcome of the analysis of the annotated data. However, our manual annotation relies on extensive pre-processing, in particular the systematic pre-marking of selected features, in an approach inspired by Biber (Biber, 1988; Biber et al., 2007). Figure 1 gives an overview of the methodology developed for this experiment. The association of exhaustive NLP-generated linguistic information (pre-marked features) with human intuitions (manual annotation of structures and cues) produces rich data, opening the way for new investigations using corpus-linguistics or data driven methods. Two multi-level structures have been annotated according to this methodology within the ANNODIS project—-topical chains and enumerative structures—-, but the present paper deals solely with the latter.

In line with the approach outlined in Section 2, the annotation project described here differs in several major ways from previous discourse annotation initiatives, such as the Penn Discourse Treebank (PDTB Prasad et al., 2008), the RST (Rhetorical Structure Theory) Treebank (Carlson et al., 2003), or the Discourse Graphbank (Wolf et al., 2004). The PDTB's focus is low-level discourse structure (elementary predicate-argument relations) and it is grounded in a lexicalised approach to discourse (role of discourse connectives as predicates). Though based on different models of discourse relations, with varying views on the role of lexical connectives, the RST Discourse Treebank and the Discourse Graphbank share similar objectives, in line with a what-perspective more than a how-perspective, to return to the distinction introduced in Section 2.1. The fact that all these annotation projects use only news material, mostly from the Wall Street Journal, is also revealing of how they differ from the experiment presented here, as will be made clear in the description of our experimental framework in Section 4.2.

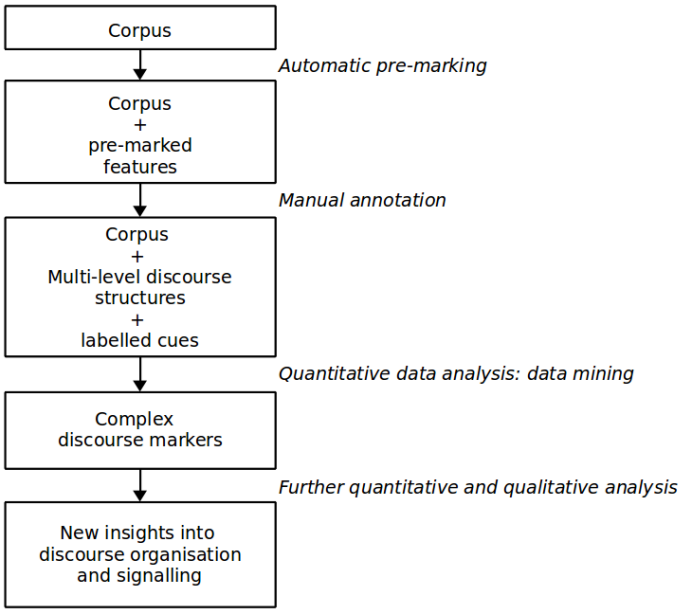



Figure 1: A data-intensive method for the study of discourse organisation and discourse signalling

### 3.1 Why enumerative structures?

Having set out the context in which the annotation experiment was designed, we will now focus on one of the two multi-level structures selected for annotation: enumerative structures beyond the sentence level, as illustrated by Example 2:

**Example 2** *from Wikipedia (English): "Global warming" (Retrieved 2014-10-09)*

*Examples of impacts include:*

- *Food: Crop production will probably be negatively affected in low latitude countries, while effects at northern latitudes may be positive or negative. Global warming of around 4.6 °C relative to pre-industrial levels could pose a large risk to global and regional food security.*
- *Health: Generally impacts will be more negative than positive. Impacts include: the effects of extreme weather, leading to injury and loss of life; and indirect effects, such as undernutrition brought on by crop failures.*

Our interest in enumerative structures as textual patterns deployed at different discourse organisation levels is initially rooted in systemic functional linguistics: Halliday's description of text as "the unit of the semantic process" (Halliday, 1977/1983, p. 63) encourages the formulation of hypotheses on how perception of high-level structures may influence text interpretation at a more local level. In this context, we are developing an approach to "texture" which takes into account visual aspects of text construction, aspects considered by Power et al. (2003) as part of "document structure". Along the lines defined by these authors—-pursuing Nunberg's reflection on "text grammar" (Nunberg, 1990)—-and by researchers inspired by Virbel's model for "text architecture" (Luc

et al., 2000; Luc and Virbel, 2001), we argue for a linguistic status for what Power et al. (2003) call "the graphical component", on a par with lexico-syntactic cues. After Luc and Virbel (2001), we describe enumerative structures as textual objects resulting from a textual act whereby text is arranged (visually or through other devices) so that the reader becomes aware of this textual arrangement. The associated semantics is that the reader is led to interpret the enumerated elements (i.e. the items) as similar in some respect, and therefore as constituting a segment homogeneous in terms of a "co-enumerability criterion". The co-enumerability criterion may be lexically expressed, as in Example 2 (*Examples of impacts*), or realised more indirectly. Two peripheral elements may contribute to this textual arrangement: a trigger which announces the enumeration and/or a closure. Enumerating appears thus as a very basic way of organising text, and a generic one in the sense that it can be resorted to for a wide range of semantic or rhetorical functions.

Despite this basic character, enumerating as a text construction strategy has not elicited much interest among discourse linguists: the "relative neglect" noted by Schiffrin in 1994, and described by her as "a surprising oversight" (Schiffrin, 1994, p. 378) still seems to apply. There are, on the other hand, quite a few studies focusing on specific linguistic elements playing a role in enumerating, in particular lexical item introducers, which have been variously named "linear integration markers" (Turco and Coltier, 1988; Jackiewicz, 2005), "sequencers" (Hempel and Degand, 2008) and "serial markers" (Bras and Schnedecker, 2013). Mostly concerned with the semantic description and classification of such markers (numerical: *firstly*, etc.; temporal: *subsequently*, *finally*, etc.; spatial: *in the first place*, etc.), these studies tend to leave out non-lexical cues such as visual devices and document structure. Research on "enumerable" nouns (Tadros, 1985, p. 6) or "shell nouns" (Francis, 1994; Schmid, 2000), so-called because of their underspecified meaning, also constitutes a relevant related field of investigation. Such nouns are seen as announcing ("predicting" to use Tadros' term) subsequent specification in the following text, and, in the case of enumerative structures, naming the co-enumerability criterion which provides the rationale for enumerating.

In contrast with these two groups of studies which mostly take specific markers as their starting point, our interest is in the text-structuring role of enumerating, and in the diverse ways in which these structures are signalled (Ho-Dac et al., 2010). Seen from this angle, the "markers" selected in the studies mentioned above are cues amongst others, playing a role in multiple-cue signalling devices—-complex discourse markers. The existing research, however, as well as providing numerous insights, raises a number of issues of interest to our project, issues which underlie the questions we are going to ask of our annotated data.

From our text organisation perspective, as distinct from the discourse markers perspective in which studies of item introducers were carried out, there is no reason to give special treatment to lexical markers or to distinguish them from the various other ways in which enumerating may be signalled. As mentioned earlier, we take full account of document structure and include among potentially relevant text features the visual devices which organise the text on the page, delimiting different spans of text: typographical variations, layout (indentation, line spacing, paragraph breaks, bullets, headings). As visual devices can be seen as pulling enumerative structures towards the textual component while lexico-syntactic cues seem better able to contribute to the ideational component, a central objective of this study is to examine how these various cues interact, and what variables may have an impact on these interactions.

Another underlying question concerns the nature of the relation between the enumeration proper, i.e. the items, and the peripheral elements: trigger and closure. Viewing the relationship between the co-enumerability criterion expressed in the trigger (or closure) and each of the items in the enumer-

ation in terms of taxonomy-based hypernym-hyponyms relations is clearly too narrow: expressions of co-enumerability may be used to gather together world entities, but also textual objects (*section*, *chapter*), rhetorical functions (*examples*, as in Example 2), or other forms of textual organisation such as steps in a chronology, stops in an itinerary, etc. Within a larger goal of exploring how the textual and the ideational components are enmeshed, enumerative structures deserve scrutiny for their ability to organise text and categorise content at one and the same time.

### 3.2 An annotation model for enumerative structures

Our annotation model was designed to allow a moderately open-ended annotation task, aiming to leave some leeway for possible off-model annotators' intuitions. According to this model, an enumerative structure (ES) extends beyond a single sentence[3] and is made up of three segments: (1) the trigger: an optional introductory segment, (2) the enumeration, defined as a list of at least two co-items, (3) the closure: an optional closing segment. Figure 2 gives a schematic representation of this definition of ESs. It shows how the elements entering into this linear arrangement are in fact different kinds of nested text spans linked via a number of possible short and long distance discourse relations[4].

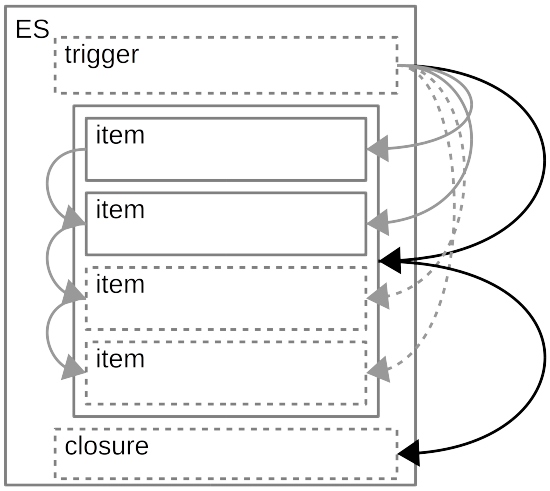



Figure 2: Enumerative structure representation

Following Luc et al. (2000), enumerating is described as a textual act which asserts the co-enumerability of the listed "entities" by transposing it textually. This "textual transposition" can take many forms, the most obvious being when items are separate paragraphs with bullet points. The nature of the co-enumerability may be made explicit, as in Example 3 below, where the two text segments are presented as similar in that they are both "criticisms" (*A moralistic criticism* [*Une critique moraliste*] and *A deterministic criticism* [*Une critique déterministe*])[5]. The co-enumerability

3. In accordance with this definition, the sentence-level ES which ends the second item of Example 2 was not included in the annotation. In some cases, however, annotators decided to include single sentence ESs.
4. The identification and annotation of these relations were outside the scope of this annotation exercise.
5. All examples from this point on are extracts from the annotated French-language corpus. For each example, the expressions which are central to the commentary are given in the text in an English translation followed by the original French (in square brackets). For complete English translations of the French-language examples, see Appendix 1.

criterion may be expressed in the trigger, in a prospective element (*two types of criticisms* [*deux types de critiques*]), and/or in the closure, in an encapsulation (*These two criticisms* [*Ces deux critiques*]) (cf. Conte, 1996; Sinclair, 1983). The enumeration is the only necessary element in this structure.

**Example 3** *from WIKI sub corpus*[6] *(wik2_liberteSE_coder3_1254325598390)*

| | | |
|---|---|---|
| *En effet, contre la liberté indépendance, il existe au moins deux types de critiques :* | *ES* | TRIGGER |
| - *Une critique moraliste : cette liberté relève de la licence, i.e. de l'abandon au désir. Or, il n'y a pas de liberté sans loi (Rousseau, Emmanuel Kant), car la liberté de tous serait en ce sens contradictoire : [...]* | | ITEM 1 |
| *On remarque que dans cette conception philosophique de la liberté, les limites ne sont pas des limites contraignant la liberté de la volonté humaine ; ces limites définissent en réalité un domaine d'action où la liberté peut exister, ce qui est tout autre chose.* | | |
| - *Une critique déterministe : s'abandonner à ses désirs, n'est-ce pas leur obéir, et dès lors un tel abandon ne relève-t-il pas d'une forme déguisée de déterminisme ? Nous serions alors victimes d'une illusion de libre arbitre : [...]* | | ITEM 2 |
| *Nietzsche reprendra cette critique : Aussi longtemps que nous ne nous sentons pas dépendre de quoi que ce soit, [...]* | | |
| *Ces deux critiques mettent en lumière plusieurs points importants. [...]* | | CLOSURE |

From Example 3 onward, the formatting of examples obeys the following conventions: the two right-hand columns delimit each annotated ES and its components; horizontal lines in the left-hand column indicate paragraph breaks in the original, i.e. each boxed segment corresponds to a complete paragraph. In Example 3 for instance, the trigger is in a paragraph, the items consist of two paragraphs each, and the closure starts a sixth paragraph. Where excessively long paragraphs were cut, this is signalled by [...] (items 1 and 2). When a component covers only part of a paragraph, this is indicated as in Example 3 for the closure.

The reference associated with each example (e.g. *wik2_liberteSE_coder3_1254325598390*) is its identifier in the ANNODIS resource. It can be searched for in the resource using the ANNODIS browser[7].

## 4. Annotating enumerative structures: annotation procedure and experimental framework

Biber et al. (2007) propose a step-by-step method for corpus-based studies of discourse, providing a detailed account of seven steps seen as necessary in order to arrive at generalisable descriptions of discourse structure in corpora. These steps may be carried out in two possible orders: either top-down (*a priori* communicative/functional categories provide the basis for manual text segmentation) or bottom-up (starting with automatic segmentation based on lexical cohesion). In both cases, the segmentation stage leads to a linguistic characterisation based on the analysis of the distribution of textual features, according to the methodology initially set up by Biber to produce an emergent text typology (Biber, 1988). In the bottom-up approach, the communicative/functional categories are derived from the linguistic characterisation (identification of clusters, which are then given a functional interpretation). Our approach is fundamentally grounded in Biber's use of systematic feature-marking and analysis, but can be seen as proposing a third way with respect to Biber et al.

6. The composition of our corpus is described in Section 4.2.2 below.

7. `http://redac.univ-tlse2.fr/corpus/annodis/me_download/ANNODIS_SE.xml`

(2007). In accordance with the top-down approach, the functional units under study are determined and defined *a priori*. This is the model which is embodied in the annotation manual, and sketched in Section 3.2 above. But in a perspective akin to Biber et al's bottom-up approach, the human annotation process is guided by the pre-marking of features which emerge from previous studies as potentially relevant for the identification and description of enumerative structures.

The next section outlines these two major steps in the annotation procedure. Section 4.2 then fills in the detail of how they were carried out in practice: it describes the annotation interface, the corpus and the annotation model.

### 4.1 A two-step annotation procedure

#### 4.1.1 AUTOMATIC PRE-MARKING OF FEATURES

Prior to manual annotation, a systematic pre-marking of potentially relevant features was automatically carried out on the POS-tagged and syntactically parsed texts[8], relying on local grammars and making use of specifically designed lexicons. The selection of features calls upon previous research (see Section 3.1) and covers a wide variety of linguistic phenomena, both visual (punctuation, layout) and lexico-syntactic. The set of pre-marked features is organised in Table 1 below into seven types, which constitute the basis for the analyses which will be presented in Section 6.

| **Feature type** | **Feature Description** |
|---|---|
| Typography and layout | indentation, line space, paragraph breaks, bullets and numbering, headings |
| Punctuation patterns | [:] preceding [;] or [,]<br>[:] in final paragraph position |
| Sequencers (sentence-initial—s.i.) | *First*, *Secondly*, *On the other hand*, *The third X* |
| Circumstance adverbials (s.i.) | including lexemes expressing time, place, categories: *Since 1956*, *In Austria*, *In linguistics*, |
| Prospective elements | specific NP patterns including specific lexemes: *the following elements*, *two types of criticisms* |
| Encapsulations | pre-verbal demonstrative NPs with a numeral as determiner: *These two criticisms*, |
| Connectives (s.i.) | *Moreover*, *To sum up*, *In contrast* |

Table 1: The set of pre-marked features

The inclusion of sentence-initial circumstance adverbials in this set of features is based on Charolles' framing hypothesis (Charolles et al., 2005): such adverbials have the potential to project forward an interpretation criterion, and thus define the initial boundary of a discourse frame, i.e. a text segment clustering around a specific interpretation criterion. They were pre-marked as potential item introducers, as were sentence-initial connectives, which are potential sequencers. The automatic detection of such sentence-initial cues proceeded in two steps: the first was to detect all syntactically detached elements occurring before the grammatical subject; the second to attribute

8. The POS-tagging was performed by Treetagger (Schmid, 1995), and the parsing by Syntex (Bourigault et al., 2005; Bourigault, 2007).

to each detached element a syntactico-semantic function (circumstantial adverbial, sequencer, other connective).

Prospective elements consist in simple and fairly unambiguous cataphoric patterns:

```
XXX as follows (.:)
PREP the following NUMBER XXXs.
```

In addition to these cataphoric patterns, prospective elements also include plural noun phrases where a classifier, or "shell-noun" (cf. Section 3.1), occurs[9]. As for encapsulations, two patterns associated with shell-nouns were used (Conte, 1996; Schmid, 2000): plural demonstrative noun phrases and noun phrases introduced by the semi-determiner *such* (*tel(le)s*).

Example 4 reproduces (3) with all pre-marked features shown in bold (prospective elements, encapsulations and punctuation). Bullets and other layout features are not highlighted, as they are, by definition, visible.

**Example 4** *from WIKI sub corpus (wik2_liberteSE_coder3_1254325598390)*

| | | |
|---|---|---|
| *En effet, contre la liberté indépendance, il existe au moins **deux types de critiques :*** | *ES* | TRIGGER |
| - *Une critique moraliste : cette liberté relève de la licence, i.e. de l'abandon au désir. Or, il n'y a pas de liberté sans loi (Rousseau, Emmanuel Kant), car la liberté de tous serait en ce sens contradictoire : [...]* | | ITEM 1 |
| *On remarque que dans cette conception philosophique de la liberté, les limites ne sont pas des limites contraignant la liberté de la volonté humaine ; ces limites définissent en réalité un domaine d'action où la liberté peut exister, ce qui est tout autre chose.* | | |
| - *Une critique déterministe : s'abandonner à ses désirs, n'est-ce pas leur obéir, et dès lors un tel abandon ne relève-t-il pas d'une forme déguisée de déterminisme ? Nous serions alors victimes d'une illusion de libre arbitre : [...]* | | ITEM 2 |
| *Nietzsche reprendra cette critique : Aussi longtemps que nous ne nous sentons pas dépendre de quoi que ce soit, [...]* | | |
| ***Ces deux critiques** mettent en lumière plusieurs points importants. [...]* | | CLOSURE |

### 4.1.2 MANUAL ANNOTATION OF STRUCTURES AND CUES

Pre-marked features were meant to act as flags to guide annotators in the identification of sporadic discourse units, leading them away from linear reading towards a more global view of text. The manual annotation consisted of two main tasks: first, delimiting and labelling the components of ESs which were detected (trigger, co-items, closure and the co-enumerability criterion if explicitly stated); second, marking-up features considered as relevant cues, by either validating pre-marked features or annotating and labelling complementary cues. In (3) for example, after delimiting the components and identifying the co-enumerability criterion (*criticisms* [*critiques*] expressed in trigger and closure), the annotators validated all the pre-marked features and went on to annotate as an extra cue the parallelism between the two NPs (*a moralistic criticism* / *a deterministic criticism*) introducing the items. Once identified, additional cues were labelled according to the categories defined in the annotation guidelines, i.e. the categories used for premarking with the addition of syntactic parallelism[10].

9. A list of 64 classifiers was manually adapted from Schmid (2000) to French.
10. This lexico-syntactic feature was not pre-marked as its automatic identification is not yet operational due to the high computational complexity of the task.

If no predefined label fitted, the annotators were invited to create descriptive labels. As a consequence, a proportion of annotator-added cues form a heterogeneous set of non-categorised features (e.g. coreferential expression, trigger repetition, apposition, named entity...).

Each feature marked-up as relevant (whether or not it was pre-marked) becomes what we call an "ES-cue" , i.e. a linguistic feature which, in combination with others, participates in the signalling of enumerative structures, and hence of discourse organisation. It is through the identification of recurring configurations of ES-cues that we propose to define complex markers (cf. Section 6).

## 4.2 Experimental framework

### 4.2.1 THE ANNOTATION INTERFACE

With annotators having to annotate textual zones of varying sizes, as well as deal with discontinuity and possible overlaps with previously delimited zones, the annotation task was highly complex and required an efficient purpose-built annotation interface.

The design of the GLOZZ interface (Widlöcher and Mathet, 2012) reflects two major requirements concerning text visualisation and the annotation procedure itself. The text visualisation interface has to take into account the output of the pre-marking procedure, including XML encoding of the text layout and formatting. The annotation interface must offer a panel of user-friendly editing tools for delimiting and characterising units; it must also facilitate navigation in the text being annotated. The solution adopted is to offer two views of the text: a main view for annotation and a global view to get a large-scale vision of the text (cf. Glozz premarked documents in Figure 3). Through these two modes of access to text, the interface encourages the annotator to combine a top-down and a bottom-up approach during reading, so as to be able to see local cues as well as global structures.

In order to ensure that the grasp of texts by annotators is as ecological as possible, and to reduce the inevitable processing difference between annotating and reading, the main view must present texts as real documents with major aspects of layout preserved.

All these requirements were taken on board in the design of the GLOZZ annotation platform[11].

### 4.2.2 THE ANNODIS CORPUS

Our approach to discourse imposes constraints on the selection of texts for the corpus. Contrary to previous discourse annotation programmes (cf. Section 2), we opted for lengthy expository texts, first because they tend not to be structured around a major referent—-as is often the case in narratives—-, secondly because they favour complex organisation and are therefore more likely to contain different structuring modes (including complex document structure). Another consideration was that corpus linguistics methods require fairly large volumes of texts and sufficient numbers of annotated structures if some generalisation of observations is to be possible (cf. Piérard and Bestgen, 2006). Finally, because we consider genre as a feature to be taken into account in the definition of complex discourse markers (Ho-Dac and Péry-Woodley, 2009; Taboada and Das, 2013), we compiled a diversified corpus enabling contrastive analyses.

Considering all these criteria, the texts selected for inclusion in the ANNODIS corpus combine three genres of lengthy expository texts: web-encyclopaedia articles (nearly 200,000 words from the French Wikipedia in its version of June 18, 2008), scientific papers (proceedings of *Congrès Mondial de Linguistique Française* 2008, about 135,000 words) and reports in the field of interna-

11. http://glozz.free.fr/

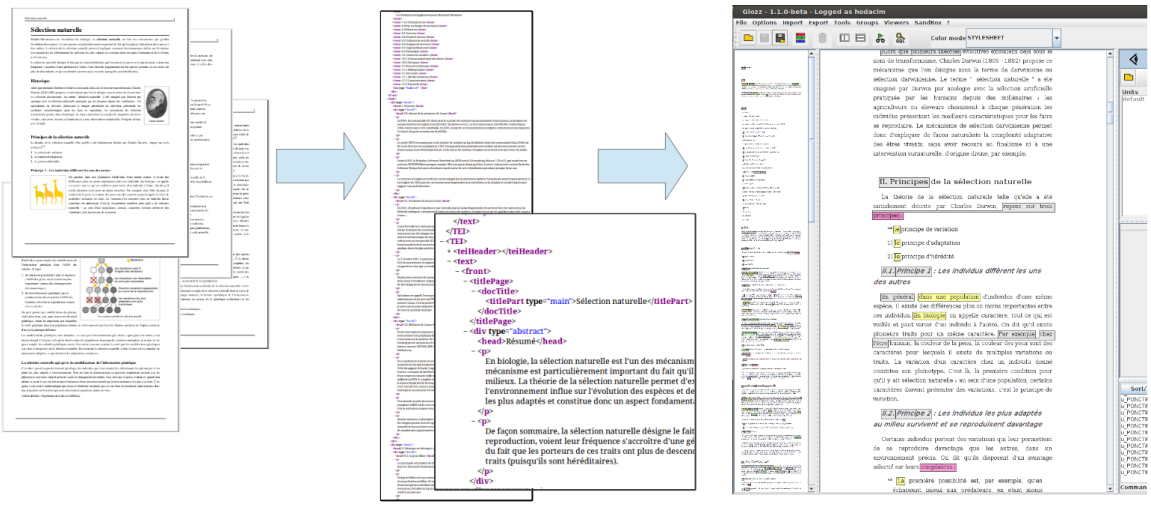



Figure 3: ANNODIS corpus preparation: from original text to pre-marked document ready for annotation

tional relations (from the *French Institute for International Relations*, over 180,000 words). These three sub-corpora are respectively named WIKI, LING and GEOP. The total number of texts (83) was set in accordance with the time constraints on the annotation programme.

Given our objectives, special care was taken in the preparation of the corpus: not only were all the texts XML encoded in conformity with the TEI-P5 norm, but it was imperative that the visual appearance of the texts be preserved, which is also a departure from previous experiments. Semi-automatic procedures were set up to annotate and encode the specific layouts signalling textual objects: title, headings with their level, paragraphs, lists and citations. Figure 3 gives a schematic view of the corpus preparation process.

### 4.2.3 THE ANNOTATION EXERCISE

The manual annotation was programmed in two stages, beginning with an exploratory phase dedicated to the evaluation of the task's feasibility, which led to a series of improvements in the procedure: clarification of the protocol, simplification of the annotation model, changes in the visualisation parameters, and correction of the annotation guide. Then came the annotation task itself. Three undergraduate linguistics students were selected as neutral (non-specialist) annotators. The 83 texts of our corpus were split into 3 sets. There was a training phase during which four texts were jointly annotated and the three annotators were encouraged to compare and share their annotations. This training phase led to an improved, stable version of the annotation guide (Colléter et al., 2012). After training, six new texts were annotated by the three coders, and these annotations were used for measuring inter-annotator agreement.

Measuring inter-annotator agreement in this case means checking whether two annotators have identified the same ESs in the target text. We considered that there was agreement on a given ES when annotators A and B had selected the same text span, and identified exactly the same items

within this span. If two annotated ESs differed only in terms of trigger and/or closure (these units being optional) while respecting the previous conditions, they were considered identical.

Overall agreement between two coders for each text was measured via the F-score. The nature of the task ruled out traditional agreement measures (such as Cohen's Kappa) because ES marking is not a categorisation task. In a task such as ours, as Hripcsak and Rothschild (2005) explain, there is no access to a negative count, i.e. we cannot take into account the fact that both annotators agreed that there are no ESs in a particular span of text. For the evaluation of a marking task, the F-score is the measure which is most commonly used (see e.g. Brants (2000) for syntactic annotation). In our case the F-score is based on the number of ESs identified by both annotators and the overall number of ESs identified by each, as formulated below:

$$F = \frac{(2 \times \#\ of\ conjointESs)}{\#ofcoderA\ ESs + \#ofcoderB\ ESs}$$

This score was measured for every pair of annotators over the 6 texts (2 from each subcorpus), each having been annotated by three different coders. The overall average F-score is 0.67 (sd 0.21), meaning that over two ESs out of three marked up by one coder were also marked up by the other coder. This value was considered sufficient for the final annotation phase to be launched, whereby the remaining texts were distributed among the three annotators, each text being dealt with only once.

In a final stage, disagreements were post-annotated, and adjudicated versions of the ten multi-annotated texts from the training and evaluation phases were produced[12]. As observed in Colléter et al. (2012), disagreements mostly concern small and/or isolated ESs, as well as structures which may be considered as contrasts or chronologies rather than ESs.

The data collected at each stage is available on-line (original documents, texts prepared for annotation, pre-adjudicated versions, etc.)[13], together with a technical report which includes the annotation manual together with coders' testimonies and adjudication details (Colléter et al., 2012). The exploitation of the annotations has so far been carried out in two ways: manually, by means of an exploration interface[14], and automatically, via data mining techniques.

## 5. Analysing the annotated corpus: enumerative structures (ESs) as a basic strategy

The rich annotated data resulting from the annotation exercise just described can now be examined for answers to the issues and questions raised in Sections 1 and 2. We start with a descriptive survey of the frequency, length and distribution of ESs in the corpus, which provides the basis for a quantitative assessment of their importance as a text construction strategy (Section 5.1), and a structural characterisation in terms of cardinality (number of items) and composition (presence/absence of a trigger and a closure) (Section 5.2). In Section 5.3 we compare bottom-up and multi-level approaches, taking advantage of the annotation of ESs in terms of discourse relations in a sub-corpus. We finally delve deeper into characteristics which are directly relevant to two major discourse organisation issues: enumerative structures are multi-level structures, capable of organising textual material at any level of granularity from entire sections to the sub-sentential level (see the typology in Section 5.4); enumerative structures are text-segmenting patterns as well as content-structuring

12. A reflection on the annotation exercise is presented in Ho-Dac and Péry-Woodley (2014).
13. http://redac.univ-tlse2.fr/corpus/annodis/me_download/index_en.html
14. http://redac.univ-tlse2.fr/corpus/annodis/me_download/ANNODIS_me_browser.html

categorisation devices, highlighting the interweaving of the textual and ideational components (Section 6).

### 5.1 Frequency, length and distribution of annotated ESs

Table 2 summarises the results of a first survey of the annotations, showing ESs to be a basic strategy frequently resorted to by writers in different genres of expository texts. There is an average of 12 ESs per text in our corpus (range: 2 to 34), and ESs have an average length of 429 words, with considerable variation (from 8 to 8,666 words). The "text coverage" value is the proportion of a given text appearing in at least one ES[15]: on average, 44.6% of a text's words are contained in ESs; in some cases, text coverage is over 90%. ESs are present in all three sub-corpora with significant variations which will be presented in the next section together with variations regarding composition.

| Sub-corpus | Texts, *n* | ESs, *n* | Mean *n* of ESs per text | Mean length (words / ES) | Text coverage (%) |
|---|---|---|---|---|---|
| WIKI | 28 | 401 | 14 | 455 | 55.5 |
| LING | 25 | 297 | 12 | 452 | 46.8 |
| GEOP | 30 | 293 | 10 | 369 | 32.8 |
| Total | 83 | 991 | 12 | 429 | 44.6 |

Table 2: Frequency and coverage of annotated ESs in ANNODIS and all three sub-corpora

The next sub-sections aim to flesh out this initial picture via analyses of the composition ESs and of their interaction with discourse relations and document structure.

### 5.2 Composition of annotated ESs

Table 3, an overall view of the composition of ESs in the corpus, shows that only a small proportion is complete with respect to the canonical three-part model—-trigger, items, closure[16].

| Sub-corpus | TRIGGER | | ITEMS | CLOSURE | | COMPLETENESS | |
|---|---|---|---|---|---|---|---|
| | *n* | % | Cardinality (mean *n* of items) | *n* | % | Complete ESs (%) | Minimalist ESs (%) |
| WIKI | 300 | 74.8 | 4.1 | 36 | 9.0 | 7.5 | 23.7 |
| LING | 230 | 77.4 | 2.9 | 46 | 15.5 | 12.1 | 19.2 |
| GEOP | 209 | 71.3 | 2.9 | 49 | 16.7 | 12.3 | 24.2 |
| Total | 739 | 74.6 | 3.4 | 131 | 13.2 | 10.3 | 22.5 |

Table 3: ES composition in ANNODIS and all three sub-corpora

In Example 5, the completeness of the structure combines with a profusion of signalling devices (in bold).

15. As ESs can be nested, a portion of text can be contained in several ESs. This phenomenon, though fairly frequent, was not taken it into account at this stage of the analysis.
16. Complete ESs have both a trigger and a closure. Minimalist ESs have neither trigger nor closure.

**Example 5** *from LING sub-corpus (ling_kleiberSE_coder3_1254143156093)*

| | | |
|---|---|---|
| ***2.2 Deux manières de nier la polysémie*** *Une réponse possible est [...]la polysémie en tant qu'association de plusieurs sens à une même forme lexicale se trouve niée de* ***deux manières*** *apparemment paradoxales* ***:*** | *ES* | TRIGGER |
| ***-a- D'une part,*** *les vocables donnés comme polysémiques se voient en quelque sorte "monosémisés" par [...]* | | ITEM 1 |
| ***-b- D'autre part,*** *de façon tout à fait inverse aux tentatives de monosémisation, on fait proliférer les sens [...]* | | ITEM 2 |
| ***Les positions -a- et -b-*** *ne sont qu'apparemment paradoxales : il n'y a aucune contradiction, d'un côté, à [...]* | | CLOSURE |

The trigger is clearly marked: it consists of the heading (*2.2. Two ways of denying polysemy* [*Deux manières de nier la polysémie*]) and the sentence following it. The prospective element in the heading, made obvious by a numeral determiner, *two ways* [*deux manières*], is reiterated in the first sentence. The items are then signalled in four complementary ways: each one makes up a paragraph, they are introduced by a dash, sequentially labelled with letters, and start with a correlative adverbial which stresses the parallelism between the two assertions (*On the one hand – On the other hand* [*D'une part – D'autre part*]). The closure ends the enumeration with an encapsulating noun phrase (*Positions -a- and -b-* [*Les positions -a- et -b-*]). This example shows how different ES-cues can reinforce one another, giving the ES high visibility.

Although Table 3 gives the average number of items per ES as 3.4, it is worth noting that 42% of ESs contain only two items, whilst rare extreme cases may comprise up to 48 items[17]. Cardinality (number of items) and length (number of words) are positively correlated, though at a marginal level (`r=0.14`). Closures are rare (13.2%), whereas most ESs start with a trigger (74.6%). Given that either trigger or closure can express the co-enumerability criterion, trigger-less ESs could have been thought more likely to have a closure, but cross-tabulation of the data does not confirm this hypothesis: only 3% of trigger-less ESs have a closure. To end this general picture, complete ESs are fairly rare (around 10%) while over 22% of ESs are minimalist, i.e. composed only of items. No significant correlation has been established between completeness and length or cardinality.

Looking at Tables 2 and 3, interesting variations across our sub-corpora begin to emerge. The largest ESs, both in length and cardinality, are found in encyclopaedia articles (WIKI), where they also cover a larger part of the text than in other sub-corpora (455 words/ES; 55.5% of total text). At the other end of the spectrum, international relation reports (GEOP) contain fewer and shorter ESs (369 words/ES) which cover a much smaller proportion of the text's surface (32.8% of total text). These variations across sub-corpora are all statistically significant ($p < 0.001$, Kruskal-Wallis test), but a clearer understanding of the text structuring role of ESs is needed to assess their linguistic significance. This is what the next two sections work towards by bringing into the picture two distinct sets of annotation, discourse relations and document structure, in order to arrive at a better characterisation of ESs' discourse function.

### 5.3 Interaction between discourse relations and ESs

As mentioned in Section 2, the ANNODIS project also involved a bottom-up annotation of discourse relations, and a part of the ANNODIS resource, labelled "ANNODIS_duo", was annotated with both ESs and discourse relations. The model and method for the annotation of discourse relations originate in Segmented Discourse Representation Theory (SDRT): coders started by segmenting

17. Only 4 ESs are made up of more than 15 items.

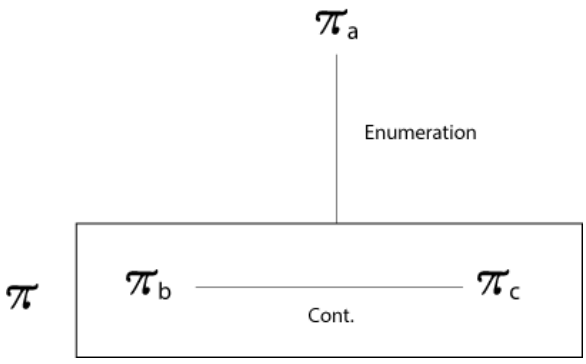
texts into Elementary Discourse Units (EDUs) and, after reaching mutual agreement, associated them with discourse relations, building up Complex Discourse Units (CDUs) until they arrived at a complete hierarchical representation of the text. Sixteen discourse relations were annotated[18], a selection which represents a compromise between informativeness and reliability of the annotation process. The selection constitutes a consensual set of relations which are shared by most discourse models, or correspond to well-defined subgroups in fine-grained theories (Hovy, 1990), as well as to the level of grain adopted for the Penn Discourse Tree Bank (Prasad et al., 2008)[19].
ESs annotated in the ANNODIS_duo resource[20] contain an average of 24.5 EDUs per ES (between 3 and 65). Because CDUs are recursively nested, the raw number of CDUs per ES is not relevant without a more qualitative analysis. Looking at the discourse relations annotated on the borders of triggers and items, i.e. relations associated to EDUs starting or ending triggers and items, the following associations may be observed:

- 75% of triggers end with an EDU linked forward to another segment via ELABORATION*[21] and/or FRAME relations,
- 94% of items start with an EDU attached to another segment via an ELABORATION* relation associated in 35% of cases with a simultaneous CONTINUITY relation,
- when considering only initial items, 92% of ESs have an initial item where the starting EDU is associated to an ELABORATION* relation.

The fact that most ESs can be described in terms of just two discourse relations, i.e. ELABORATION* and CONTINUITY, confirms that the structure can legitimately be regarded as a functional unit, regardless of the diverse forms in which it occurs. Moreover, each of these relations seems to have a specific role in the structure: ELABORATION* between the trigger and items, and CONTINUITY between items, as Example 5 illustrates. These observations also support the SDRT model developed in Bras et al. (2008) and Vergez-Couret et al. (2008) for explaining long distance attachments between trigger segments and CDUs introduced by pairs of discourse markers such as *d'abord/ensuite* (*first/then*). As a consequence, a new relation called ENUMERATION was introduced by Vergez-Couret et al. (2012) as defined in Figure 4.

According to its authors, the ENUMERATION relation was introduced so that analysts would be able "to juggle between constituents describing semantic content and constituents describing discourse packaging while ENTITY-ELABORATION would not have allowed so" (Vergez-Couret et al., 2011). The authors clearly sense that two different types of text-building process are at play, which they want to account for while keeping them apart, hence the "juggling". This term calls to mind the doubts expressed in Goutsos (1996, p. 257) as to the possibility of dealing with essentially textual relations in ideational terms:

18. EXPLANATION, GOAL, RESULT, PARALLEL, CONTRAST, CONTINUATION, ALTERNATION, ATTRIBUTION, BACKGROUND, FLASHBACK, FRAME, TEMPORAL-LOCATION, ELABORATION, ENTITY-ELABORATION, COMMENT
19. Like the annotation guide for multi-level structures, the guide produced for the annotation of discourse relations is freely available (Muller et al., 2012). It provides an intuitive introduction to discourse segments, including the question of embedding of discourse segments to form CDUs; a list of detailed instructions describing how to handle segmentation; and a semantic definition of each discourse relation with examples and potential markers.
20. 26 ESs: 15 in WIKI, 5 LING and 6 in GEOP.
21. ENTITY-ELABORATION and ELABORATION relations are merged for this analysis under the label ELABORATION*.




*Co-items ($\pi_b, \pi_c$) together introduce a complex constituent ($\pi$) which is attached to the trigger ($\pi_a$) by the* ENUMERATION *relation. A coordinating relation (by default* CONTINUATION*) is inferred between the co-items.*

Figure 4: Discourse structure for a classical enumerative structure (Vergez-Couret et al., 2012)

> "Ideational analyses of texts have identified relations of Joint, List or Sequence (Hoey, 1979; Mann and Thompson, 1988), whose status is clearly not so prominently ideational as textual. More generally, it is doubtful whether essentially presentational relationships like enumeration or listing can be couched in ideational terms at all. [...] the insistence on recognising a semantic relation between every single text segment comes into conflict with the occurrence of purely descriptive, propositionally loosely or arbitrarily related chunks of text."

A similar question was raised in an earlier study on enumerating by Luc et al. (1999), who argued that their initial representation within Virbel's Text Architecture Model (cf. Section 2.1) needed complementing by a Rhetorical Structure Theory representation, while pointing out the inadequacy of tree-like structures to represent ESs and stressing the importance of visual clues, largely overlooked by researchers working within RST. Regarding the latter argument, Virbel et al. (2005, p.234) denounce linguistics' blindness to visual cues :

> "Linguistics, just as—-to a lesser extent—-information science, has long been 'blind' to the role of visual properties of written language, while other research fields (anthropology, history of texts, cognitive and experimental psychology) did point to the fundamental importance of these properties from the viewpoint of cognition."[22].

Among the visual properties overlooked by linguists, including discourse linguists, are titles and headings, whose role in text processing has been the object of much study in cognitive psychology (Lorch and Lorch, 1996; Lemarié et al., 2008, 2012), but which are difficult to integrate within a discourse relations approach. The present study was designed from the outset to deal with the corpus not just as texts but as documents, whose layout structure is meaningful. The next section

22. "Les sciences du langage comme, dans une moindre mesure, celles de l'information sont restées longtemps "aveugles" au rôle des propriétés visuelles du langage inscrit, alors que d'autres recherches (anthropologie, histoire des textes, psychologie cognitive et expérimentale) avaient signalé l'importance fondamentale de ces aspects du point de vue cognitif."

focuses on an annotation layer dedicated to the documents' layout structure, which appears to be particularly well-suited to characterising the annotated ESs in their diversity.

### 5.4 Enumerative structures are multi-level: a granularity-based typology

As mentioned in Section 4.2.2, the layout structure of documents was annotated according to TEI-P5 encoding. The textual objects considered here are sections, headings, lists and paragraphs. They are used as features in order to account for the variety of annotated ESs in terms of length and composition. Because they enter into hierarchical relationships, these layout units also provide a scale for describing ESs' granularity level (cf. Section 2.4).

We observed earlier that in terms of completeness, ESs show variations that are not explained either statistically by length or cardinality (cf. Sections 5.1 and 5.2) or by distinct discourse relations (cf. Section 5.3). In contrast, granularity level appeared as the most informative variable for the classification of these structures (see Ho-Dac et al., 2010). The interaction between ESs and the document's layout structure, presented in Table 4, provides the basis for a granularity-based typology of ESs. This granularity-based typology emerged as the optimal way of clustering the annotated ESs according to quantifiable variations in their form and composition. Moreover, it gives us a way of organising the data by distinguishing classes of objects likely to make use of different signalling modes, as described in Section 6.

Each type is now defined more precisely in terms of its typographical and layout features.

- Type 1 corresponds to multi-section ESs, in which items are sections with a visible heading, as in Example 6 below.
- Type 2 ESs are prototypical formatted lists where each item is signalled by a bullet or number, as in Examples 2 and 3 above.
- ESs which extend over more than one paragraph but do not belong to either of the previous types are Type 3. These Type 3 ESs contain at least one paragraph break which may occur between two components (e.g. between trigger and first item or, as in Example 8 between final item and closure as well as between items), with no specific constraints on the position and/or number of paragraph breaks.
- Finally, Type 4 stands for ESs contained within a single paragraph, see Example 10 and 11. It is the most frequent across the whole corpus.

| ES type | Description | Nb of ESs | | Mean length |
|---|---|---|---|---|
| | | Nb | % | (Nb of words per ES) |
| Type 1 | multisection | 126 | 12.7 | 1,858 |
| Type 2 | bulleted list | 244 | 24.6 | 184 |
| Type 3 | multiparagraph | 216 | 21.8 | 449 |
| Type 4 | (intra)paragraph | 405 | 40.9 | 120 |
| All Types | | 991 | 100 | 429 |

Table 4: Granularity-based typology of ESs

There is considerable variation in the distribution of types across sub-corpora, as shown in Figure 5. WIKI ESs are the most strongly associated with visual layout: in 19% of cases, the items

are headed sections (Type 1) and in 36.4% they are formatted lists (Type 2). Such emphasis on visual properties is to be expected in texts designed to be read on screen. It is in marked contrast with linguistics papers and international relations reports, where ESs aligned on visual layout are fairly rare (fewer than 10% of Type 1 ESs) and Type 4 ESs i.e. low-level structures without visual properties, are the most frequent (45.5% in LING and 61.4% in GEOP against 22.4% in WIKI). Type 3 ESs, multi-level structures without headings or bullets as item introducers, are the most stable across corpora (between 20% and 23%). These variations support our assumption that genre must be considered a relevant feature for the characterisation ES and also for the definition of complex discourse markers. For example, (sub)headings may be considered as ES-cues only in specific genres or text types.

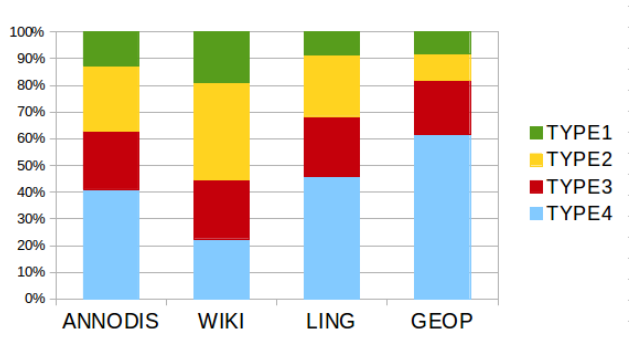



Figure 5: Granularity-based ES types across sub-corpora

As mentioned above, the proposed typology emerged as the best way to explain the variation in the form and composition of ESs. An overview of how ES composition varies across types is given in Table 5, followed by descriptions of the four ES Types.

| ES Type<br>Sub-corpus | TRIGGER<br>% | ITEMS<br>Cardinality (mean nb of items) | CLOSURE<br>% | COMPLETENESS<br>Complete ESs (%) | <br>Minimalist ESs (%) |
|---|---|---|---|---|---|
| Type 1 | 84.1 | 3.4 | 4.0 | 4.0 | 15.9 |
| Type 2 | 96.3 | 4.1 | 13.1 | 13.2 | 3.7 |
| Type 3 | 60.6 | 3.7 | 20.8 | 16.2 | 34.7 |
| Type 4 | 65.9 | 2.8 | 12.1 | 7.4 | 29.3 |
| All Types | 74.6 | 3.4 | 13.2 | 10.3 | 22.5 |

Table 5: Typology of ESs - composition of annotated ESs

### 5.4.1 TYPE 1: MULTISECTION ESS

Type 1 ESs cover several sections of a document, with section headings signalling co-items. The least frequent in all three sub-corpora, they are, as can be expected, the longest ESs in our corpus, averaging 1,858 words (with an enormous range: 252 to 8,666 words), yet their cardinality is close to the average (3.4 items per ES). The few Type 1 ESs which have a closure (4%) are all complete

ESs i.e. they also have a trigger. Indeed, most Type 1 ESs have a trigger (84%) which is generally a heading of the next level up and announces the enumeration both visually (via document structure) and semantically. Example 6 shows a Type 1 ES with a trigger and 2 items:

**Example 6** *from WIKI sub-corpus (wik2_julesCesarSE_coder2_1254907327695)*

| ***6. Les conquêtes amoureuses de César*** | *ES* | TRIGGER |
|---|---|---|
| ***6.1. Les femmes de la haute société romaine***<br>*D'après l'historien latin Suétone, César séduit de nombreuses femmes tout au long de sa vie et plus particulièrement celles issues de la haute société romaine.* | | ITEM 1 |
| *Il aurait ainsi séduit Postumia, la femme de Servius Sulpicius, Lollia, [...]* | | |
| *César entretient des relations particulières avec Servilia Caepionis, [...]* | | |
| *Le penchant de César pour les plaisirs de l'amour semble également attesté par [...]* | | |
| ***6.2. Les reines***<br>*César a des relations amoureuses avec Eunoé, femme de Bogud, roi de Mauritanie.* | | ITEM 2 |
| *Cependant, sa relation avec Cléopâtre VII est restée plus célèbre. [...]* | | |

The relation between a section heading and its sub-headings could arguably be seen as an inclusion relation similar to that of co-items in an enumerative structure. Yet it is not the case that all headed sections including headed sub-sections can be classed as ESs, and indeed most were not identified as ESs by the annotators. What marks out the annotated Type 1 ESs is the presence of a semantic criterion linking the items, in other words the fact that they function on both the ideational level and the textual level: in Example 6, the first level heading, *Cæsar's amorous conquests*, provides this semantic criterion which unites under the category *Cæsar's amorous conquests* upper-class Roman women (*Les femmes de la haute société romaine*) and queens (*Les reines*).

### 5.4.2 TYPE 2: BULLETED LISTS

Type 2 ESs are characterised by the presence of bullets or numbers signalling each item. They have the highest cardinality (4.1 items/ES), but are significantly shorter (184 words/ES, $p < 0.001$), their constituent items being generally restricted to short phrases, as in Example 7 below. There are exceptions, however, such as Example 3 above, where some items cover several paragraphs. Triggers are almost systematically present: 95% in WIKI, 97% in GEOP, and 100% in LING. The corollary is a tiny percentage of minimalist ESs.

**Example 7** *from WIKI sub-corpus (wik2_telecommunicationsSE_coder2_1255513359128)*

| ***Parmi** les principaux organismes de normalisation-standardisation mondiaux, **citons :*** | *ES* | TRIGGER |
|---|---|---|
| - *l'ETSI : European Telecommunication Standards Institute ou Institut européen des normes de télécommunication ;* | | ITEM 1 |
| - *l'ITU : International Telecommunication Union ou Union internationale des télécommunications ;* | | ITEM 2 |
| - *l'IETF : Internet Engineering Task Force ;* | | ITEM 3 |
| - *l'ATM Forum ;* | | ITEM 4 |
| - *l'ANSI : American National Standard Institute ;* | | ITEM 5 |
| - *l'IEEE : Institute of Electrical and Electronics Engineers.* | | ITEM 6 |

### 5.4.3 TYPE 3: MULTIPARAGRAPH ESS

Type 3 ESs stretch over at least two paragraphs, with no headings or bullets, as illustrated in Example 8. This example is a case of nesting: ES2 is embedded in ES1. The larger ES (ES1) is Type 3, with two paragraph breaks: one between the two items, the other before the closure. This example illustrates the role of paragraph-initial position in the signalling of ESs: each item-paragraph

starts with a sequencer (*A first observation* [*Une première observation*]; *A second observation* [*Une deuxième observation*]). These sequencers are echoed by the two item-introducing expressions in the embedded Type 4 ES (ES2) (*In the first case* [*Dans le premier cas*]; *The second position* [*La seconde position*]).

**Example 8** *from LING sub-corpus (ling_kleiberSE_coder3_1254142826046)*

<table>
<tr><td>Une première observation est à faire à ce niveau. On constate dans l'abondante littérature sur la multiplicité des [...]</td><td>ES1</td><td>ITEM 1</td><td></td><td></td></tr>
<tr><td>Une deuxième observation concerne le niveau où s'exerce la critique du fait polysémique.<br>Si on part de la conjonction définitionnelle provisoire - i - et - ii -, la polysémie peut être remise en cause, soit en critiquant - i -, soit en critiquant - ii -.<br>Dans le premier cas, celui où - i - est faux, mais où - ii - subsiste, les relations de - ii - sont à porter au crédit de la construction [...]<br>La seconde position, celle où l'on conserve - i -, mais où l'on refuse - ii -, revient à transformer un cas de polysémie en un cas d'homonymie. Elle est, c'est significatif, beaucoup moins [...]</td><td></td><td>ITEM 2</td><td>ES2</td><td>TRIGGER<br>ITEM 1<br>ITEM 2</td></tr>
<tr><td>Nos deux observations tirent dans la même direction : elles montrent que c'est avant tout le point - i -, celui de [...]<br>[...]</td><td></td><td>CLOSURE</td><td></td><td></td></tr>
</table>

Precise alignment of elements (trigger, items, closure) with paragraphs is not mandatory for this type. Example 9 shows a complete enumerative structure where the trigger and the first two items share one paragraph, while the last item and the closure appear in a separate one.

**Example 9** *from GEOP sub-corpus (geop_19SE_coder1_1253605625609)*

<table>
<tr><td>Par ailleurs, une guerre contre l'Irak pouvait se faire selon trois scénarios.</td><td>ES</td><td>TRIGGER</td></tr>
<tr><td>Le premier consistait à renouveler l'expérience de 1991 (sans doute avec une coalition amoindrie); il nécessitait des mois de préparation et posait de réels problèmes de politique intérieure aux Etats-Unis.</td><td></td><td>ITEM 1</td></tr>
<tr><td>Le deuxième scénario consistait à répéter l'expérience de décembre 1998, à savoir [...]</td><td></td><td>ITEM 2</td></tr>
<tr><td>Le troisième scénario consistait à envoyer sur place plusieurs commandos de services spéciaux chargés de liquider le dictateur [...]</td><td></td><td>ITEM 3</td></tr>
<tr><td>De fait, la première option semblait être la seule permettant de poursuivre l'effort de 1991, en poussant [...]</td><td></td><td>CLOSURE</td></tr>
</table>

Type 3 ESs are average in length, with slightly above average cardinality. Whilst the frequency of triggers is markedly low (61%), closures are much more frequent than elsewhere, particularly in LING (27%) and GEOP (32%). Despite this comparatively high frequency of closures, Type 3 includes the highest proportion of minimalist ESs: over a third have neither trigger nor closure. These minimalist Type 3 ESs are characterised by the presence of series of ES-cues in paragraph-initial position (see Section 5.2.2 above). It may also be noted that only in Types 3 and 4 do we find ESs which have a closure and no trigger, as ES1 in Example 8.

### 5.4.4 TYPE 4: INTRAPARAGRAPH ESS

Type 4 ESs, which are contained within a paragraph, are the most frequent. The ES in (10) below and the nested structure (ES2) in (8) are examples of this type. Unsurprisingly, Type 4 ESs have the smallest mean length (120 words/ES); they also have significantly fewer items than other types: over half are 2-item ESs (against 29% for Type 2, 34% for Type 1 and 41% for Type 3). The presence of triggers and closures is slightly below average. As a consequence, complete ESs are fairly rare (7%), in contrast with minimalist ESs (29%), illustrated in Example 10.

**Example 10** *from GEOP sub-corpus (geop_11SE_coder1_1254301361468)*

| | | |
|---|---|---|
| *[...]* | | |
| ***Entre 1949 et 1970,*** *la part de la demande couverte par le pétrole importé est passée de 10 % à 23 %.* | *ES* | ITEM 1 |
| ***Entre 1978 et 1985,*** *les importations ont fortement baissé, tant en valeur absolue (- 3,8 Mb / j) que relative (- 16 points de part de marché). Deux facteurs expliquent ce phénomène : le développement du champ géant de Prudhoe Bay en Alaska, et la chute de la demande pétrolière liée au second "choc pétrolier" de 1979 et à la récession économique.* | | ITEM 2 |
| ***À partir de 1985,*** *la part du pétrole importé dans la couverture de la demande n'a cessé d'augmenter, jusqu'à aujourd'hui.* | | ITEM 3 |

To summarise, the major variations accounted for by the granularity-based typology are as follows:

- Type 1 ESs are significantly longer;
- Type 2 ESs, with higher cardinality and shorter length, have a trigger most of the time and as a consequence are rarely minimalist ESs;
- Type 3 ESs have significantly more often a closure, but are also more minimalist than the others;
- Type 4 ESs are the shortest in length and cardinality with a high proportion of minimalist ESs.

This typology will be used in the next section to organise the data by distinguishing classes of objects likely to make use of different signalling modes.

## 6. Mining the annotated corpus for configurations of ES-cues

In this final section, we move on to the search for recurring configurations of ES-cues as a way of identifying the complex markers signalling ESs. Prior to this phase of the analysis, ES-cues (i.e. validated features) had to be organised into relevant categories. We added syntactic parallelism, encountered in various forms in Examples 1, 4 and 5, as a frequent annotator-added cue which had been identified from the outset as a potential ES-cue but could not be pre-marked for technical reasons. We now describe this re-classification, which accounts for the differences between Table 1 (Section 3) and Table 6 below.

In addition to the main annotation task, identifying ESs and their constitutive elements, the annotators were asked to mark up the cues which they identified as signalling these structures (cf. 3.1.2). The resulting corpus contains 4,052 individual annotated cues which were explicitly identified as ES-cues either through pre-marked feature validation or through manual addition; to these must be added 500 headings, systematically counted as ES-cues when occurring in a trigger or when item-initial.

It should be stressed that identifying cues is considerably more difficult than identifying ESs, and at this stage we have no inter-annotator agreement measure on this task. A number of problems were encountered, some of which originate in the pre-marking procedure—-any text processing program inevitably generates both noise and silence—-, others in the level of linguistic competence required of the annotators, or in semantic difficulties inherent in some of the cues. Due to these limitations, our analysis will be restricted to the identification of the global behaviour of ES-cues.

The goal of this section is twofold:

1. to examine frequencies and distributions for the different kinds of ES-cues;
2. to identify recurrent cue configurations as a first step towards the definition of ES markers (cf. Section 3.1).

Table 6 below lists the categories of ES-cues taken into account. The abbreviations in bold are used throughout the remainder of this section.

| Nb of cues | Description | Example |
|---|---|---|
| in TRIGGER | | |
| 443 | **TriggerLex.**: prospective elements and other lexical features | Ex. 3 (*deux types de critique*), 5 and 6 |
| 302 | **TriggerPunct.**: punctuation patterns | Ex. 3, 5 and 7 |
| in ITEM | | |
| 595 | **Paral.**: syntactic parallelisms | Ex. 3 (item-initial NPs) |
| 628 | **Seq.**: sequencers and connectives | Ex. 5 and 8 |
| 649 | **Adv.**: circumstance adverbials | Ex. 10 |
| 433 | **ItemHead.**: headings at the beginning of items | Ex. 6 |
| 1065 | **Bullets**: bullets and numbering | Ex. 3 |
| 246 | **ItemPunct.**: punctuation patterns | |
| 88 | **ItemOthers:** other lexical features | |
| in CLOSURE | | |
| 103 | **ClosureLex.**: encapsulations, connectives and other lexical features | Ex. 3: *Ces deux critiques*<br>Ex. 9: *To sum up* / *En somme* |

Table 6: Categories of ES-cues for analysis (after re-classification)

Annotator-added cues, except syntactic parallelisms, were counted as "TriggerLex.", "ClosureLex." or "ItemOthers" according to their host component. We are aware that these categories are excessively broad. We will in particular need to isolate prospective elements and encapsulations in order to investigate the expression of the co-enumerability criterion. A semantic characterisation of the expression of the co-enumerability criterion is required for a finer functional classification of ESs.

### 6.1 Description of cues in ES components

Tables 7 and 8 provide the detail of the distribution of annotated cues for each component. Distributions are given both globally and according to ES types. All values are percentages, and are relative to the frequency of the corresponding element: out of the 131 ESs with a closure (i.e. out of 13.2% of ESs, cf. Table 5) 78.6% have a lexical cue. Percentages do not add up to 100 for triggers and items, as they each can have between zero and several cues (of different kinds).

#### 6.1.1 TRIGGER AND CLOSURE CUES

Trigger and closure are almost systematically signalled by a cue: over 75% of these components were associated by the annotators with at least one ES-cue.

Two categories vary considerably in frequency across types: explicit lexical elements, which potentially announce the co-enumerability criterion (TriggerLex.), and punctuation marks (a final

| | | Trigger | | | Closure | |
|---|---|---|---|---|---|---|
| | | | % with cue | | | |
| ES Type | (Nb.) | Nb. | TriggerLex. | TriggerPunct. | Nb. | % with ClosureLex. |
| Type 1 | (126) | 106 | 41.5 | 3.8 | 5 | 80.0 |
| Type 2 | (244) | 235 | 55.3 | 76.6 | 32 | 68.8 |
| Type 3 | (216) | 131 | 74.8 | 19.8 | 45 | 75.6 |
| Type 4 | (405) | 267 | 64.0 | 34.5 | 49 | 87.8 |
| All Types | (991) | 739 | 59.9 | 40.9 | 131 | 78.6 |

Table 7: Distribution of trigger and closure cues

colon, TriggerPunct.), which have a purely textual role. Characteristic trigger punctuation is most frequent in Type 2 ESs, part of a well-established pattern for introducing lists, seen in Examples 3, 5 and 7. Punctuation cues are also fairly frequent in Type 4 ESs: Example 11 illustrates how punctuation is instrumental in signalling such intraparagraph ESs, with a colon as a trigger cue, and final commas reinforcing the parallelism between items.

**Example 11** *from GEOP sub-corpus (geop_27SE_coder2_1282829750411)*

*[...]*
*Les phénomènes terroristes prolifèrent au croisement de* ***quatre grandes circulations :***
***celle des*** *mots et des images (qui permet de bricoler des solidarités entre des sociétés très différentes),*
***celle des*** *capitaux (qui autorise la mise sur pied de logistiques performantes),*
***celle des*** *armes (qui ouvre sans cesse le champ des dangers futurs),*
***et celle des*** *hommes.*
*[...]*

| *ES* | TRIGGER |
|---|---|
| | ITEM 1 |
| | ITEM 2 |
| | ITEM 3 |
| | ITEM 4 |

Closures are characterised by the strong presence of lexical cues (ClosureLex.). It must however be kept in mind that this component is rare in our annotated data (cf. Table 3), with the consequence that these percentages correspond to very few cases and the results cannot be extrapolated.

### 6.1.2 ITEM CUES

Table 8 shows that all predefined categories of item cues were indeed found in the ANNODIS resource, and that conversely very few item cues are found in the "ItemOthers" miscellaneous category. Over 15% of ESs contain at least one of the most common lexical cues i.e. a sequencer, an adverbial or a parallelism. Among these lexical cues, only parallelisms are distributed more or less equally in all ES types. As a consequence, parallelism is the cue which combines most frequently with the visual cues inherent to Type 1 and 2 ESs (ItemHead. or Bullets). The other lexical cues are on the contrary extremely rare in Type 1 and Type 2. This lack of variety in the signalling of items creates a stark contrast between Types 1 and 2 on the one hand, and Types 3 and 4 on the other, the latter displaying a greater complexity of organisation associated with a wide range of cues.

The distribution of item cues in Types 3 and 4 presents an interesting contrast: Type 3 ESs favour circumstance adverbials over sequencers (47.8% and 26.3% respectively), whereas in Type 4 sequencers (34.4%) prevail over adverbials (19.3%). An explanation for this difference may be found in the organising role of circumstance adverbials (cf. 1.2): in order to function as discourse segmentation markers, these must be paragraph-initial, as in Example 12 below where 3 place adverbials occupy paragraph-initial position (*In the United States [Aux États-unis*], *In Germany [En Allemagne*], *In Spain [En Espagne]*.

| ES Type | Nb of Items | % with ItemHead. | Bullets | ItemPunct. | Seq. | Adv. | Paral. | ItemOthers |
|---|---|---|---|---|---|---|---|---|
| Type 1 | 434 | 100.0 | 0.5 | 0.0 | 1.6 | 6.0 | 17.1 | 4.1 |
| Type 2 | 995 | 0 | 100.0 | NA | 2.0 | 2.1 | 16.8 | 2.4 |
| Type 3 | 802 | 0 | 7.2 | 2.6 | 26.3 | 47.8 | 15.6 | 4.2 |
| Type 4 | 1134 | 0 | 5.3 | 16.8 | 34.4 | 19.3 | 20.2 | 1.1 |
| All Types | 3365 | 13.1 | 31.1 | 7.3 | 18.7 | 19.3 | 17.7 | 2.6 |

Table 8: Distribution of item cues (cue category per ES type)

**Example 12** *from WIKI sub-corpus (wik2_attentats11septSE_coder1_1254125810843)*

| | | |
|---|---|---|
| ***Aux États-Unis,*** *la seule personne à avoir été jugée jusqu'à présent pour son implication directe avec les attentats du 11 Septembre est le Français Zacarias Moussaoui. Arrłté moins d'un mois avant les attaques, il a été accusé par les autorités fédérales américaines d'avoir eu connaissance des attentats à venir mais de n'avoir pas communiqué ses informations. Le 3 mai 2006, au terme de deux mois de procès, il a été reconnu coupable par le jury du tribunal fédéral d'Alexandria en Virginie de six chefs d'accusation de complot en liaison avec les attentats terroristes du 11 Septembre et condamné à la prison à perpétuité, sans possibilité de remise de peine.* | *ES* | ITEM 1 |
| ***En Allemagne,*** *le marocain Mounir al-Motassadeq arrłté le 28 novembre 2001, est condamné une première fois à quinze ans de prison en 2003 pour complicité dans ces attaques. Remis en liberté en février 2006 après que sa condamnation a été cassée, il voit sa première peine confirmée par le tribunal de Hambourg le 8 janvier 2007.* | | ITEM 2 |
| ***En Espagne,*** *le Syrien Imad Eddin Barakat Yarkas, chef de la cellule locale d'Al-Qaida est arrłté le 13 novembre 2001, inculpé de conspiration en vue des attentats de septembre 2001. Il est condamné le 26 septembre 2005 à vingt-sept ans de prison.* | | ITEM 3 |

This positional constraint is incompatible, by definition, with intraparagraph Type 4 ESs. Sequencers on the other hand are specialised in the signalling of ESs, and seem therefore to be more independent from positional constraints, which could explain why they are particularly suited to these low-level ESs, as illustrated by Example 8.

In addition, punctuation item cues are fairly frequent in Type 4, as are punctuation trigger cues. Example 13 illustrates such a combination of punctuation and lexical cues in a Type 4 ES where items are separated by a semicolon and the last one introduced by the connective *enfin* / *finally*.

**Example 13** *from GEOP sub-corpus (geop_16SE_coder1_1255425907703)*

| | | |
|---|---|---|
| *[...]* | | |
| *Mais les discussions sont occultées par des positions idéologiques : 'la subvention est intrinsèquement néfaste', 'la PAC est intouchable', 'les PED sont quoi qu'il arrive victimes d'un système injuste' ... positions contredites par les pratiques.* | *ES* | TRIGGER |
| *Tout le monde subventionne, même les pays les plus vertueux, d'une façon qui peut fausser les échanges* ***;*** | | ITEM 1 |
| *la PAC est en constante révision, et son coût n'est pas élevé (0,5 % du PIB européen)* ***;*** | | ITEM 2 |
| ***enfin****, il est faux que les PED aient tout à gagner d'une disparition totale des subventions, tant est grand l'avantage comparatif des plus gros producteurs agricoles, qui ne sont pas des PED.* | | ITEM 3 |

### 6.2 Cue associations

The examples make it clear that most ESs are signalled concurrently by several kinds of ES-cues. The previous section gave an insight into the frequencies of individual ES-cues without taking into

account their co-occurrence. Yet our hunch, as stressed in Section 2, is that textual patterns are not just signalled by discrete clearly identifiable dedicated markers, but by configurations of ES-cues functioning as complex discourse markers. In such a perspective, a discourse function should not be attributed to a particular lexical expression—-on the basis of a specific semantic or pragmatic value—-but rather to this expression when it occurs in a particular context or configuration (cf. the pattern formed by the series of paragraph-initial adverbials in Example 1). The ANNODIS resource now provides us with data to investigate this hypothesis, and this is what we attempt below using the notion of cueset. This is not an easy task, however, as we want to allow for flexibility while hoping to catch recurring patterns. The identification of textual patterns when reading can be conceptualised in terms of pattern recognition: a threshold is reached when there are enough converging cues to push interpretation towards the identification of the pattern in question. A more satisfactory approach to *cuesets* would involve attributing weights to individual ES-cues in order to account for the fact that several weak cues may do the same work as one strong cue. This is our horizon, with the work presented here as a first exploration of the data in this direction.

A *cueset* is the set of *cue categories* occurring in an ES. As the purpose of these sets is to help identify frequent cue associations, we apply the following simplifications:

1. for item cues, a single occurrence suffices for the cue to be included in the set, there is no need for the cue to appear in every item;

2. cue frequency within an ES is not taken into account, and is reduced to a simple binary value of presence/absence.

The main reasons for these simplifications are the potential incompleteness of item marking (e.g. *firstly* in the first item not followed by other sequencers), and the inherent difficulty of the cue-annotation task.
The number of different associations was calculated for the whole collection of ESs, and for each ES type studied independently. Of all theoretically possible configurations[23], over half were actually observed: 113 distinct cuesets were identified for all 991 ESs. This result is interpreted as meaning, on the one hand, that ESs are signalled by a variety of cue configurations—-for example a lexical cue in the trigger followed by a series of sequencers, or a combination of sequencers and adverbials; and on the other hand, that certain specific cue associations recur, while others are not found.

In order to identify the most frequent cuesets, we focus here on the 14 cuesets occurring at least 20 times across types and corpora, which represent 63% of all ESs. Among the most frequent cuesets, we find clusters made of cues which have not been the focus of much attention in studies of enumerating i.e bullets, punctuational patterns and more interestingly lexical cues in the trigger (cf. Example 7). In almost all frequent cuesets there is at least one trigger cue (a punctuational or a lexical one) associated with all possible item cues (i.e. headings, bullets, sequencers, adverbials, parallelisms).

The most frequent cueset is the combination $TriggerPunct. + TriggerLex. + Bullets$ which occur 83 times i.e. in 8.4% of ESs. The fairly similar cueset $TriggerPunct. + Bullets$ recur only 40 times, which means that visual cues are usually combined with lexical ones. The same kind of combination is found with the cueset $TriggerPunct. + TriggerLex. + ItemPunct.$ This finding supports our view of signalling as a struggle between different forces (cf. Section 2.2):

23. There are 210 theoretically possible configurations since all 10 ES-cues listed in Table 6 may signal ESs.

visual devices can be seen as pulling enumerative structures towards the textual component while lexico-syntactic cues seem better able to contribute to the ideational component.

Cuesets including the much-studied sequencers are also very frequent. Two types were observed with approximately equal frequency: cuesets composed of sequencers only as in Example 8 (73 cuesets, 7.4% of ESs); and cuesets which combine sequencers with other cues such as a lexical cue in the trigger (as in Example 9), parallelisms or adverbials (87 cuesets, 8.8% of ESs). Cuesets made up purely of adverbials (Examples 9 and 11) are as frequent as those made up purely of sequencers (74 cuesets, 7.5% of ESs). But cuesets mixing adverbials with other kind of cues are fairly rare (only 23 with lexical cue in the trigger and 20 with sequencers).

All the cuesets described are fairly stable across sub-corpora except for two: those made up of ItemHead. and Paral. which only recur in scientific papers and those made up of adverbials which occur primarily in encyclopaedia articles and never in scientific papers.

A number of specific configurations have been identified, which, when correlated with ES types, can be summarised as follows:

- Type 1 ESs are typically signalled by a sequence of same level headings, with an upper level heading acting as a trigger, and occurs in documents where layout and visual formatting play a prominent role.
- Type 2 ESs are typically signalled by a sequence of bulleted items, almost systematically introduced by a punctuational cue (final colon in the preceding paragraph), and/or a lexical cue in the trigger which may carry semantic information on the co-enumerability criterion.
- Type 3 ESs are typically signalled by a contiguous series of paragraphs with circumstance adverbials (or, less likely, sequencers) in initial position, which have both a textual and an ideational role. Such structures seem to be highly genre-sensitive.
- Type 4 ESs can be described as a single paragraph containing a series of sequencers, with a high probability of a colon marking the end of the trigger, or a prospective element indicating the co-enumerability criterion. In contrast with Type 2, Type 4 ESs reflect the ideational dimension of discourse organisation more than the textual dimension.

### 6.3 Towards complex discourse organisation markers identification

In order to validate and formalise the cuesets observed above, we used a common data mining technique for identifying recurrent associations between pairs of cues by extracting the association rules i.e. the logical implication rules between cues (Agrawal et al., 1993). This method ensures that all possible cases are systematically examined. The association rules are of the form:

> *X1 & X2 ... Xn → Y*
> *(where Xi and Y are cue types), meaning that most (at least 75%) of ESs having X1, X2 and Xn as cues also have Y.*

This technique identified the following rules (in order of decreasing systematicity):

1. Bullets → TriggerPunct.
2. ItemPunct.→ TriggerPunct.
3. TriggerPunct. & Bullets → TriggerLex.

4. ItemHead.→ TriggerLex.

5. ClosureLex. → TriggerLex.

Rules 1 and 2 indicate that whenever the items are bulleted (Type 2 ES) or have a punctuation mark (essentially Type 4 ESs), they have a trigger with a punctuation cue (colon), and vice-versa. This reflects the coherence of patterns of punctuation marks in ESs.

Rules 3 and 4 merely confirm that ESs of Types 1 and 2 have a high proportion of triggers (Table 5), and that most of these triggers contain a lexical cue (Table 7). Such a finding can be interpreted as showing that even in apparently purely visual i.e. textual ESs, a lexical cue somewhere will ensure the presence of the ideational dimension.

Rule 5 links the existence of an encapsulation to that of a prospective element. Again, this must be interpreted with the knowledge that both these cues are quite systematic in triggers and closures. Rule 5 says that ESs with a closure generally have a trigger. In other words, closures are not used to compensate for the absence of a trigger. By lowering the tolerance of the association rules system, more rules can be made to emerge, although they are known to be much less reliable and systematic. One interesting point is that, even with a low threshold, no rule involving circumstance adverbials emerges. This negative result confirms that this cue category is much less likely to work in association with others.

Most of these results were predicted and explained in previous sections, which suggests that no other obvious specific cue associations can be identified as a result of our annotation exercise.

## 7. Conclusion

Enumerative structures were selected as the focus of this study as a way of throwing new light on linearisation and segmentation, discourse phenomena which are particularly difficult to analyse empirically. We described enumerative structures in Sections 2 and 3 as a generic multi-level device for organising text. According to our broad functional definition, they are textual patterns assembling text spans which are made to appear as similar in a given respect, thereby forming a higher-level segment homogeneous in this particular respect. They arrange into linear format text segments which are ideationally discontinuous but functionally equivalent and interchangeable. Their signalling calls upon a great diversity of cues working together. As such, enumerative structures constitute good handles for analysing how writers cope with "The Unbearable Linearity of Texts".

Our objective of proposing a data-intensive methodology for the study of linearisation and segmentation imposed certain requirements in terms of ease of detection and annotation. As their function depends on their being readily detectable, they constitute a good object for annotation. The sizeable annotated resource described here, which has been made available to the research community, is characterised by a number of original features:

- it is composed of highly-structured long expository texts in three different genres (as opposed to short news material);

- its mark-up combines NLP-based exhaustive techniques and human intuitions;

- the visual characteristics of the texts have been encoded so as to provide a presentation respectful of the original layout;

- the annotation guidelines were designed so as to bring together under a functional umbrella objects that are linguistically more diverse than in previous studies: Type 4 (intraparagraph ESs) is shown to be just one realisation, accounting for no more than 40% of annotated ESs.

The paper summarises the results of the first analyses of the annotated data. The evidence-based and quantified typology we propose encapsulates the major results of our analyses so far: it provides a broad picture of a device realising a basic textual strategy. The preliminary analyses presented here only scratch the surface of what the annotated corpus allows. Where cues have been counted together, e.g. in the analysis of triggers and closures, finer analyses are needed to take into account the specific contribution of each type of cue, in particular in the case of expressions of the co-enumerability criterion. Qualitative studies are necessary for the analysis of the rhetorical and semantic functions of enumerative structures in text, opening the way for the study of correlations between such functions and cue configurations, and for the exploration of the differences between sub-corpora.

One important issue concerns the nature of the relation between the items and the "classifier" which introduces and links them. Does an enumerative structure reveal a pre-existing categorisation or can it "discursively create" such knowledge, as suggested by Schiffrin (1994, p. 396) or Luc et al. (2000)? The latter hypothesis, a constructivist one more in keeping with our textual approach, makes the expression of the co-enumerability criterion worth studying as potentially revealing not just of pre-existing knowledge structures, but of a writer's discourse strategy. A systematic study of expressions of the co-enumerability criterion is under way (Rebeyrolle and Péry-Woodley, 2014). It suggests that only in very few cases are these expressions linked to the enumerated items by a hypernym-hyponym relation, i.e. a taxonomic relation which is discourse-independent ($< 10\%$). First results show that the expressions of the co-enumerability criterion would generally be better described as "text-bound" labels (Francis, 1994), in terms of shell nouns or signalling nouns (Flowerdew, 2003; Flowerdew and Forest, 2015). The study develops a model associating semantic and textual properties of linguistic expressions of co-enumerability, proposing that semantic characteristics situate enumerative structures containing these expressions on a cline from mainly textual (metadiscursive) to mainly ideational (stable, discourse-independent categorisation). On this basis, a classification of ESs in terms of their discourse function will be put forward, to be compared with existing taxonomies of relevant discourse relations and analyses based on them (e.g. Joint, List, Sequence in RST).

Cue configurations should also be examined further in relation to layout (ES type) and composition: for example minimalist ESs (neither trigger nor closure) are markedly more numerous in Types 3 and 4, which is also where adverbials are most frequent as item markers, and may compensate for the absence of expression of the co-enumerability criterion in a prospective or encapsulating element (Rebeyrolle and Péry-Woodley, 2014). We wish to look further into these trade-offs as examples of how ideational and textual metafunctions are interwoven in these structures. ESs should also be examined in context, within the linearity of text: interactions between ESs (nested and in sequence), interactions between ESs and other textual structures (including annotated topical chains). Finally, the markers identified should be tested and refined for the automatic detection of ESs, with potential applications in automatic text synthesis and document navigation.

## References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages –, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `https://hal.archives-ouvertes.fr/hal-00976087`.

Rakesh Agrawal, Imielinski Tomasz, and Arun Swami. Mining association rules between sets of items in large databases. *ACM SIGMOND Record*, 22(2):207216, 1993.

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

Nicholas Asher, Farah Benamara, Myriam Bras, Lydia-Mai Ho-Dac, and Philippe Muller. Annodis and related projects: case studies on the annotation of discourse structure. In Nancy Ide and James Pustejovsky, editors, *The Handbook of Linguistic Annotation*. Springer, Berlin, to appear.

John Bateman, Thomas Kamps, Jörg Kleinz, and Klaus Reichenberger. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3): 409–449, 2001.

Douglas Biber. *Variation across speech and writing*. Cambridge University Press: Cambridge, Massachusset, 1988.

Douglas Biber, Ulla Connor, and Thomas A. Upton. *Discourse on the move, using corpus analysis to describe discourse structure*, volume 28 of *Studies in corpus Linguistics*. John Benjamins Publishing Company: Amsterdam/Philadelphia, 2007.

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

Didier Bourigault, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques, and Silvia Ozdowska. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, 2005.

Thorsten Brants. Inter-annotator agreement for a german newspaper corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000.

Myriam Bras and Catherine Schnedecker. Dans un (premier+second+nième) temps vs en (premier+second+nième) lieu : qu'est ce qui fait la différence? *Langue Française*, 179:89–108, 2013.

Myriam Bras, Laurent Prévot, and Marianne Vergez-Couret. Quelles relations de discours pour les structures énumératives? In Bernard Laks Jacques Durand, Benot Habert, editor, *Congrès*

*Mondial de Linguistique Française - CMLF'08*, pages 1945–1964. EDP Sciences, Institut de Linguistique Française, Paris, 2008.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.

Wallace L. Chafe. *Subject and Topic*, chapter Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View, pages 25–55. New York/San Francisco/London: Academic Press, 1976.

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

Michel Charolles, Anne Le Draoulec, Marie-Paule Péry-Woodley, and Laure Sarda. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies*, 15(2):203–218, 2005.

Maud Colléter, Cécile Fabre, Ho-Dac Lydia-Mai, Marie-Paule Péry-Woodley, Josette Rebeyrolle, and Ludovic Tanguy. La ressource annodis multi-échelle : guide d'annotation et bonus. *Carnets de grammaires*, 20:1–63, 2012. URL `https://hal.archives-ouvertes.fr/hal-00983076`.

Marie-Elisabeth Conte. Anaphoric encapsulation. *Belgian Journal of Linguistics: Coherence & Anaphora*, 10:1–9, 1996.

Nils Erik Enkvist. A parametic view of word order. In E.Szer, editor, *Text Connexity Text Coherence: Aspects Methods Results*, pages 320–336. Helmut Buske: Hamburg, 1985.

John Flowerdew. Signalling nouns in discourse. *English for specific purposes*, 22(4):329–346, 2003.

John Flowerdew and Richard W Forest. *Signalling Nouns in Academic English*. Cambridge University Press, 2015.

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

Morton Ann Gernsbacher. The structure building framework: What it is, what it might also be, and why. In B. K. Britton and A. C. Graesser, editors, *Models of text understanding*, pages 289–311. Lawrence Erlbaum Associates: Mahwah, New Jersey, 1995.

Morton Ann Gernsbacher. Two decades of structure building. *Discourse processes*, 23(3):265–304, 1997.

Dionysis Goutsos. *Modeling Discourse Topic: sequential relations and strategies in expository text*, volume 59. Greenwood Publishing Group, 1997.

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

Michael A.K. Halliday. Text as semantic choice in social contexts. In J. Webster, editor, *The Collected Works of M.A.K. Halliday (Volume 2): Linguistic Studies of Text and Discourse*, page 2381. London: Continuum, 1977/1983.

Hilde Hasselgård. *Adjunct adverbials in English*. Cambridge: Cambridge University Press, 2010.

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

Laurent Heurley. Processing units in written texts: paragraphs or information blocks? In J. Costermans and M. Fayol, editors, *Processing interclausal relationships: studies in the production and comprehension of text*, pages 179–200. Lawrence Erlbaum Associates: Mahwah, New Jersey, 1997.

Lydia-Mai Ho-Dac and Marie-Paule Péry-Woodley. A data-driven study of temporal adverbials as discourse segmentation markers. *Discours*, 4:http://discours.revues.org/5952, June 2009. doi: 10.4000/discours.5952. URL `https://hal.archives-ouvertes.fr/hal-00979739`.

Lydia-Mai Ho-Dac and Marie-Paule Péry-Woodley. Annotation des structures discursives : l'expérience annodis. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, and Sophie Prévost, editors, *4e Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 2647 – 2661, Berlin, Germany, July 2014. doi: 10.1051/shsconf/20140801286. URL `https://hal.archives-ouvertes.fr/hal-01068119`.

Lydia-Mai Ho-Dac, Marie-Paule Péry-Woodley, and Ludovic Tanguy. Anatomie des structures énumératives. In *Traitement Automatique des Langues Naturelles*, page (publication numérique), Montréal, Canada, 2010. URL `https://halshs.archives-ouvertes.fr/halshs-00509189`.

Lydia-Mai Ho-Dac, Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle, and Ludovic Tanguy. An empirical approach to the signalling of enumerative structures. *Discours*, 10:(publication en ligne), 2012. URL `https://halshs.archives-ouvertes.fr/halshs-00954182`.

Michael Hoey. Signalling in discourse. discourse analysis monographs 6. *English Language Research, Birmingham University, Birmingham*, 1979.

Eduard H. Hovy. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, pages 128–136, Dawson, PA, June 1990.

George Hripcsak and Adam S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 12(3):296–298, 2005.

Agatha Jackiewicz. Les séries linéaires dans le discours. *Langue française*, 148:95–110, 2005.

Julie Lemarié, Robert Frederick Lorch, Hélène Eyrolle, and Jacques Virbel. Sara: A text-based and reader-based theory of signaling. *Educational Psychologist*, 43(1):27–48, 2008.

Julie Lemarié, Robert Frederick Lorch, and M.-P. Péry-Woodley. Understanding how headings influence text processing. *Discours, special issue on Signalling discourse organisation – Multidisciplinary Approaches to Discourse 2010 (MAD 10)*, 10, 2012. URL `http://discours.revues.org`.

Willem J.M. Levelt. The speaker's linearization problem. *Philosophical Transactions Royal Society London*, B295:305–315, 1981.

Robert Frederick Lorch and Elizabeth Pugzles Lorch. Effects of organizational signals on free recall of expository text. *Journal of educational psychology*, 88(1):38, 1996.

Christophe Luc and Jacques Virbel. Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum*, 23(1):103–123, 2001.

Christophe Luc, Mustapha Mojahid, Jacques Virbel, Claudine Garcia-Debanc, and Marie-Paule Péry-Woodley. A linguistic approach to some parameters of layout: A study of enumerations. In *AAAI 1999 Fall Symposia "Using Layout for the Generation, Understanding or Retrieval of Documents*, pages 20–29, North Falmouth, Massachussetts, 1999.

Christophe Luc, Mustapha Mojahid, Marie-Paule Péry-Woodley, and Jacques Virbel. Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes de CIDE 2000 (Colloque International sur le Document Électronique)*, pages 21–40, 2000.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

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

Daniel Marcu. Automatic discourse parsing. In K.Brown, editor, *Encyclopedia of Language and Linguistics*, pages 649–654. Elsevier, Oxford, 2nd edition, 2006.

Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Farah Benamara, Myriam Bras, Anne Le Draoulec, and Laure Vieu. Manuel d'annotation en relations de discours du projet annodis. Technical Report 21, Carnets de grammaires, CLLE-ERSS, 2012.

Geoff Nunberg. The linguistics of punctuation. *csli, Lecture Notes, University of Chicago Press*, 18, 1990.

Marie-Paule Péry-Woodley, Stergos Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher. Le corpus annodis, un corpus enrichi d'annotations discursives. *TAL*, 52(3):71–101, 2011.

Sophie Piérard and Yves Bestgen. Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL*, 47(2):89–110, 2006.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. Document structure. *Computational Linguistics*, 2(29):211–260, 2003.

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

Josette Rebeyrolle and Marie-Paule Péry-Woodley. Énumration et structuration discursive. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, and Sophie Prévost, editors, *4e Congrès Mondial de Linguistique Française (CMLF 2014)*, pages 3183–3196, Berlin, Germany, July 2014.

Deborah Schiffrin. Making a list. *Discourse processes*, 17(3):377405, 1994.

Hans-Jörg Schmid. *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin, New York: Mouton de Gruyter, 2000.

Helmut Schmid. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28, 1995.

John Sinclair. *Planes of discourse. The Twofold Voice: Essays in Honour of Ramesh Mohan.* Salzburg: Universitt Salzburg., 1983.

Maite Taboada and Debopam Das. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse*, 4(2):249–281, 2013.

Maite Taboada and William C Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459, 2006.

Angela Tadros. *Prediction in text*. Number 10. English Language Research, 1985.

Gilbert Turco and Danielle Coltier. Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques*, 57:57–79, 1988.

Teun A van Dijk. Story comprehension: An introduction. *Poetics*, 9(1):1–21, 1980.

Marianne Vergez-Couret, Laurent Prévot, and Myriam Bras. Interleaved discourse, the case of two-step enumerative structures. In *Constraints in Discourse III*, 2008.

Marianne Vergez-Couret, Myriam Bras, Laurent Prévot, Laure Vieu, and Caroline Atallah. Discourse contribution of enumerative structures involving" pour deux raisons". In *Constraints in Discourse 2011*, page online, 2011.

Marianne Vergez-Couret, Laurent Prévot, and Myriam Bras. How different information sources interact in the interpretation of interleaved discourse: The case of two-step enumerative structures. *Discours*, 11:(publication en ligne), 2012. doi: 10.4000/discours.8743. URL `http://discours.revues.org/8743`.

Jacques Virbel. The contribution of linguistic knowledge to the interpretation of text structures. In *Structured documents*, pages 161–180. Cambridge University Press, 1989.

Jacques Virbel. Textual enumeration. In *Texts, Textual Acts and the History of Science*, pages 221–266. Springer, 2015.

Jacques Virbel, Claudine Garcia-Debanc, Thierry Baccino, Lartitia Carrio, Corinne Dominguez, Christian Jacquemin, Christophe Luc, Mustapha Mojahid, Marie-Paule Péry-Woodley, and Sabine Schmid. Approches cognitives de la spatialisation du langage. de la modélisation de structures spatiolinguistiques des textes l'expérimentation psycholinguistique : le cas d'un objet textuel, l'énumération. *Cognitique, Agir dans l'espace, Éditions de la Maison des Sciences de l'homme*, 2005.

Tuija Virtanen. *Discourse Functions of Adverbial Placement in English: Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions*. Abo Akademi University Press: Abo, 1992.

Tuija Virtanen. Point of departure: Cognitive aspects of sentence-initiale adverbials. In T. Virtanen, editor, *Approaches to cognition through text and discour*, pages 79–97. Berlin/New York: Mouton de Gruyter, 2004.

Antoine Widlöcher and Yann Mathet. The glozz platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering*, pages 171–180. ACM, 2012.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. Discourse graphbank. *Linguistic Data Consortium, Philadelphia*, 2004.

## Appendix

**NB1.** The aim of this appendix is to give readers access to the French-language examples so they can follow the arguments developed in the text. We have therefore opted for translations which remain close to the original French, sometimes at the expense of the quality of the resulting English.

**NB2.** Many of the examples given cover in reality large stretches of text and have been cut, sometimes extensively, for the sake of clarity and brevity (cuts are indicated by [...] in the text). The reference associated with each example is its identifier in the ANNODIS resource. It can be searched for in the resource using the ANNODIS browser (`http://redac.univ-tlse2.fr/corpus/annodis/me_download/ANNODIS_SE.xml`).

**Example 3 from WIKI sub corpus (wik2_liberteSE_coder3_1254325598390)**

Against (the idea of) freedom as independence, there are at least two types of criticisms:

A moralistic criticism: this freedom would be a form of licentiousness, i.e. surrender to ones desires. Now, there is no freedom without law (Rousseau, Emmanuel Kant), as freedom for everyone would be a contradiction in terms: [...]

One notices that in this philosophical conception of freedom, the limits are not limits that constrain the freedom of human will; A deterministic criticism: is surrendering to ones desires not a way to obey them? and does such a surrender not amount in the end to a hidden form of determinism? We would in this case be under an illusion of free choice: [...]

Nietzsche picks up this criticism: As long as we do not feel dependent on something, [...]

These two criticisms throw light on several important points.

[...]

**Example 5 from LING sub-corpus (ling_kleiberSE_coder3_1254143156093)**

**2.2 Two ways to negate polysemy**

A possible reply is [...]. [...] polysemy as the association of several meanings to one lexical form is negated in two apparently paradoxical ways:

-a- On the one hand, word forms given as polysemous become in a way "monosemised" by [...]
-b- On the other hand, in a manner which is quite the opposite of the monosemisation attempts, meanings are made to proliferate [...]

Positions -a- and -b- are only apparently paradoxical: there is no contradiction between, on the one hand [...]

**Example 6 from WIKI sub-corpus (wik2_julesCesarSE_coder2_1254907327695)**

## VI. Cæsar's amorous conquests

### VI.1. Women from Roman high society

According to the Roman historian Suetonius, Cæsar won over many women in the course of his life, in particular women belonging to Roman high society. It is said that he won the love of Postumia, Servius Sulpicius' wife, of Lollia, [...]

Cæsar had a special relationship with ServiliaCaepionis, [...]

Evidence of Cæsar's taste for the pleasure of love also comes from [...]

**VI.2. Queens**

Cæsar had an affair with Euno, wife of Bogud, the king of Mauritania.

However, his relationship with Cleopatra has remained most famous.

**Example 7 from WIKI sub-corpus (wik2_telecommunicationsSE_coder2_1255513359128)**

Among the major international normalisation-standardisation bodies, let us mention:

- l'ETSI: European Telecommunication Standards Institute ou Institut europen des normes de tlcommunication;
- l'ITU: International Telecommunication Union ou Union internationale des tlcommunications;
- l'IETF: Internet Engineering Task Force;
- l'ATM Forum;
- l'ANSI: American National Standard Institute;
- l'IEEE: Institute of Electrical and Electronics Engineers.

**Example 8 from LING sub-corpus (ling_kleiberSE_coder3_1254142826046)**

A first observation must be made at this stage. One notices in the abundant literature on the multiplicity of [...]

A second observation concerns the level at which the criticism of the polysemous fact is conducted. If one starts from the temporary definitional conjunction -i- and -ii-, polysemy can be questioned, either by a criticising -i- , or by criticising -ii-. In the first case, where -i- is false but -ii- subsists, the relations in -ii- can be credited to the construction [...] The second position, conserving -i- but refusing -ii-, amounts to transforming a case of polysemy into a case of homonymy. This position is, and this is significant, much less [...]

Our two observations go in the same direction: they show that it is first and foremost point -i- , the one [...]

**Example 9 from GEOP sub-corpus (geop_19SE_coder1_1253605625609)**

Besides, a war against Iraq could take place according to three scenarios. The first consisted in reproducing the 1991 campaign (probably with a reduced coalition); it required months of preparation and confronted the US with real domestic policy problems. The second scenario consisted in repeating the 1998 campaign, i.e. [...]

The third scenario consisted in sending several special services commandos there, with the order to remove the dictator [...] Indeed, the first option seemed to be the only one that made it possible to pursue the 1991 effort, while pushing [...]

**Example 10 from GEOP sub-corpus (geop_11SE_coder1_1254301361468)**

[...] Between 1949 and 1970, [...] ; [...] the share of demand covered by imported oil rose from 10% to 23%. Between 1978 and 1985, imports went down sharply, in absolute terms (-3.8MB/d) as well as in relative terms (-16 points of market share). Two factors explain this phenomenon: the development of the giant oil field at Prudhoe Bay in Alaska, and the fall in oil demand linked to the second "oil crisis" in 1979 and to the economic recession. Since 1985, the share of imported oil in covering demand has continuously increased up to today.

**Example 11 from GEOP sub-corpus (geop_27SE_coder2_1282829750411)**

[...] Terrorist events proliferate at the crossroad between four great circulations: the circulation of words and images (which makes it possible to cobble together solidarities between very different social groups), the circulation of capital (which allows the setting up of efficient logistics), the circulation of weapons (which keeps opening up the prospects for future dangers), and the circulation of men. [...]

**Example 12 from WIKI sub-corpus (wik2_attentats11septSE_coder1_1254125810843)**

In the US, the only person until now to have been judged for direct implication in the 9/11 attacks is the Frenchman Zacarias Moussaoui. Arrested less than a month before the attacks, he has been accused by the American federal authorities of having had knowledge of the forthcoming attacks and of not having communicated his information. On May 3rd 2006, after a two-month trial, he was found guilty by the jury of the federal tribunal of Alexandria in Virginia on six charges of conspiracy linked to the terrorist attacks of September 11th and sentenced to life imprisonment without the possibility of parole.

In Germany, the Morrocan Mounir al-Motassadeq, arrested on November 28th 2001, is sentenced first to 15 years in prison in 2003 for complicity in these attacks. Freed in February 2006 after his conviction was quashed, he saw his initial sentence confirmed by the tribunal of Hamburg on January 8th 2007.

In Spain, the Syrian Imad Eddin Barakat Yarkas, chief of the local Al-Qaida cell, is arrested on November 13th 2001, charged with conspiring towards the September 2001 attacks. On September 26th 2005 he receives a twenty-seven year prison sentence.

**Example 13 from GEOP sub-corpus (geop_16SE_coder1_1255425907703)**

[...] But the discussions are obscured by ideological positions: "subsidies are intrinsically bad", "you can't touch the CAP", "developing countries will always be the victims of an unfair system" ... positions which are not borne out by practices. All countries, even the most virtuous, use subsidies in a manner which may distort trade; the CAP is constantly under revision, and its cost is not high (0.5% of European GNP); finally, it is not true that developing countries stand to gain from a total end to subsidies, given the massive comparative advantage of the biggest agricultural producers, which are not developing countries.