Corpus annotation of macro discourse structures

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

CORPUS ANNOTATION OF MACRO DISCOURSE STRUCTURES

Lydia-Mai Ho-Dac, Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle
Université de Toulouse 2, France,
hodac@univ-tlse2.fr, cfabre@univ-tlse2.fr, pery@univ-tlse2.fr, rebeyrol@univ-tlse2.fr

We present our discourse annotation project, ANNODIS, which aims to make available a diversified French corpus annotated with discourse information, along with a set of tools for annotation and corpus exploitation. An original aspect of the project is that it combines two theoretically and methodologically different points of view on discourse: bottom-up and top-down. In the bottom-up perspective, basic constituents are identified and linked *via* discourse relations. In a complementary manner, the top-down approach starts from the text as a whole and focuses on the identification of configurations of cues signalling higher-level text segments, in an attempt to address the interplay of continuity and discontinuity within discourse. The focus of this paper is the annotation scheme used in the top-down approach, which revolves around *enumerative structures*. These structures, which are of particular interest to our project because of their ability to occur in nested configurations and at all levels of granularity (from within a sentence to across text sections), are the discourse object chosen to "bootstrap" our approach. We describe the different stages involved: corpus selection, pre-processing and "marking" techniques, and the specific interface facilities, designed to make it possible for coders to navigate and scan the text in order to identify relevant spans at different granularity levels.

keywords: Corpus annotation, discourse organisation, text segmentation, macro discourse structures, computational linguistics

I. A MACRO APPROACH TO DISCOURSE ORGANISATION

The work presented here is part of a larger project, the ANNODIS project, the aims of which is to create an annotated corpus of French written texts for the study of discourse organisation. Discourse organisation is difficult to study with corpus linguistics methods because of the lack of

resources in this domain. A major aim of this project is to make available to the research community a) a usable corpus of French expository texts exhibiting a wide range of discourse cues and annotated with various discourse phenomena; b) a set of computational tools for consulting corpus data. This paper focuses on one type of discourse phenomena manually annotated: macro discourse structures.

Corpus annotation in the ANNODIS project is characterised by a two-way approach to discourse organisation: bottom-up and top-down. The bottom-up approach consists in applying a compositional and logical model of discourse organisation (here Segmented Discourse Representation Theory: SDRT) on the corpus. The top-down approach, the concern of this paper, aims to build a model for annotating macro discourse structures within the framework of discourse sequentiality.

I.1. Discourse sequentiality in expository texts

The main purpose of the top-down approach, also called macro approach, is to develop a model for annotating discourse sequentiality and macro discourse structures. Discourse sequentiality is defined in Goutsos (1996) to explain how discourse is structured or, in Systemic Functional Linguistics terminology, to explain texture.

“Texture can be defined as the process whereby meaning is channelled into a digestible current of discourse 'instead of spilling out formlessly in every possible direction' (Halliday 1994:311) [...]” (Martin 2001:35)

With a clear focus on structure rather than on individual units of text, Goutsos proposes a three-tier model for discourse organisation: the cognitive level describes the writer's mental representation as structured by the basic strategies of continuity and discontinuity, while the linguistic level is concerned with the techniques available to realise these strategies, so as to produce, at the textual level, a segmentation into continuation and transition spans. The reader in turn uses knowledge of the instructional meaning of these signals to construct a mental representation. As a result, the sequentiality model describes text as a “periodic alternation of

transition and continuation spans" (Goutsos 1996: 501).

Goutsos' model is specifically designed to account for discourse organisation in expository texts, such as those making up the ANNODIS corpus, as strategies of continuity and discontinuity seem to be more complex in expository texts. Expository texts are defined here as topic oriented texts, after Berman & Nir-Sagiv (2007), whose experimental psycholinguistic study aims at describing, in a global and local approach, how children and teenagers produce texts in different genres and organise them. Topic oriented texts "focus on concepts and issues and express the unfolding ideas, claims, and arguments in terms of the logical interrelations among them" (Berman & Nir-Sagiv 2007:80), while narrative texts are agent-oriented and "express the unfolding of events in a temporal framework" (Berman & Nir-Sagiv 2007:79). As a consequence, discourse organisation in expository texts tends to be more complex, which increases the importance of explicit signalling[1]. On the basis of their experimental results, Berman & Nir-Sagiv conclude that expository texts result from a "top-down, topic-motivated global-level text construction" (Berman & Nir-Sagiv 2007:108). These studies provide support for our choice of expository texts for ANNODIS.

I.2. Discontinuity and text segmentation

In this framework of sequentiality in expository texts, a model is being developed for the manual annotation of continuation and transition spans. The annotated corpus will then make possible the study of the linguistic techniques used to realise the basic strategies of continuity and discontinuity. If continuity is the default, as Goutsos suggests, a major task in the writing process is to signal discontinuity. In the absence of a cue to the contrary, the reader will interpret incoming sentences as continuous. Our purpose in the macro approach is to locate discontinuity relations. In terms of text segmentation, this consists in delimiting the boundaries of discourse segments, as discourse segments may be defined in terms of internal continuity (relative to a specific homogeneity criteria) as well as discontinuity in relation to neighbouring segments.

1 In procedural texts, signalling discourse organisation can be absolutely vital.

Studies concerned with text segmentation offer a wide range of syntactico-semantic definitions of "discourse segment" and of potential cues for the signalling of discourse discontinuities. Most of the definitions of discourse segments are concerned with thematic organisation signalled with cues related to referential continuity and thematic breaks. For example, Hearst (1997), Piérard & Bestgen (2008) *inter alia* characterize segments in terms of thematic continuity because they define discourse organisation mostly in those terms. In contrast to this view, we posit multiple levels of discourse organisation: writers and readers have to concurrently manage several levels which include thematic organisation but also temporal and spatial organisation, rhetorical articulations and logical document structure (realised by the visual architecture of the document *e.g.* paragraph and chapter subdivisions). Moreover, these different levels of organisation may be embedded: we may for example find inside a segment characterised by thematic continuity several (sub)segments characterised by different temporal references or different points of view, as illustrated in example (1) (extract from the introduction of an international relations article).

Example (1)

Depuis la fin de la guerre froide, **le débat entre spécialistes des relations transatlantiques** s'est trop souvent contenté d'osciller entre les bons sentiments et la simplification. **Il** ne s'est pas suffisamment porté sur l'ampleur des changements de fond rendus inévitables par le changement de système international produit par l'effondrement du régime soviétique. La première tendance, parfois marquée par une frilosité nourrie par la crainte de remettre en cause l'édifice institutionnel issu de la guerre froide, s'est exprimée le plus souvent sous la forme de satisfecits donnés à l'Alliance atlantique pour ses progrès supposés en matière d'adaptation aux conditions de l'après-guerre froide, et parfois sous la forme plus dynamique de projets d'élargissement, géographique et fonctionnel de l'OTAN et de l'Union européenne. Les travaux de la Rand Corporation, et en particulier ceux de Larabbee, Asmus, Gompert et Kugler, avaient ainsi contribué en leur temps à lancer le débat sur l'élargissement de l'OTAN à trois pays qui a finalement abouti en 1999.

Plus récemment, **la discussion** s'était portée sur un éloignement supposé des valeurs sociales entre les deux rives de l'Atlantique, auquel les événements du 11 septembre 2001 ont au moins provisoirement mis fin. **Ce débat** se poursuit, mais il est maintenant limité à la sphère de l'analyse sociale. En termes de politique étrangère, **cette discussion sur la dérive des continents** a pris la forme d'une opposition entre l'unilatéralisme de la politique américaine et le multilatéralisme de leurs partenaires européens.

Des visions divergentes [heading]

Le moindre mérite de l'article de Robert Kagan, dont Commentaire a publié la version française, n'est pas de sortir **le débat** de cette ornière....

In this extract, a string of cohesion devices, typically pronominal and lexical anaphora, establishes continuity around the Theme of "*debate between specialists of transatlantic relations*" (these devices appear in bold). All these devices constitute cues helping the reader understand that the writer keeps referring to the same thing *i.e.* that there is a main continuation span constructed around a thematic continuity. Concurrently, there are two sentence-initial time adverbials (highlighted in the same way as headings) each introducing a temporal frame extending over the entire paragraph. All the sentences in each paragraph can therefore be said to cluster around a common time reference expressed by the sentence-initial time adverbial. The scope of these time adverbials constructs two (sub-)continuation spans organised around a temporal continuity. Finally, as the first two paragraphs constitute the introduction of the paper, they can be said to be homogeneous in respect of rhetorical (or argumentative) function. Figure 1 summarizes our sequentiality analyses for example (1):

Figure 1: Segmenting representation of example (1)

Enkvist depicts this multidimensional global organisation as a "struggle between the different forces that affect the linearization of discourse" (Enkvist 1985).

A consequence of this "struggle" is that discourse cues signalling sequentiality and more precisely discontinuity may be simultaneously contributing to several of these interdependent levels. This is why, when asking human coders to annotate, discontinuity appears as something corresponding to a multitude of discourse phenomena, from the more local to the more global level,

participating in thematic, circumstantial or rhetorical organisation. Unless precise guidelines are provided, annotating discontinuity will inevitably be a very subjective task. An annotation model must guide coders towards more objective discourse phenomena, while at the same time establishing weak constraints in order to cover a wide range of phenomena. Enumerative structures strike us as a good starting point. They can be seen as a *meta structure* covering a range of discourse organisation phenomena and providing a concrete illustration of a certain kind of sequentiality.

II. MODELLING SEQUENTIALITY THROUGH ENUMERATIVE STRUCTURE

II.1. Enumerative structure in the broad sense

Enumerative structures correspond to a textual strategy used to adapt discourse to the linear format of text. These structures are of great interest because they are not limited to a single dimension of discourse presentation: the enumeration principle can be instantiated into topical, temporal or rhetorical structures. They organise discourse hierarchically by enumerating different sub-topics, properties, events, processes, arguments etc. constitutive of a single hyper-Theme or macro-Theme. Luc *et al.* (2000) give the following account of this "text object":

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

Enumerative structures are realised and interpreted through interacting high- and low-level cues (*e.g.* visual cues such as bullets vs. lexical choices). They epitomise the linearity constraint, setting out a great diversity of discourse elements in the linear format of written text. A canonical enumerative structure comprises three segments:

1. a **trigger** (*i.e.* an introductory segment) announcing the enumeration, potentially containing

the enumeration's macro-Theme or a **prospective element** *i.e.* a generic cataphoric expression (*e.g.* comprises [three segments], may be explained by [the following factors]);

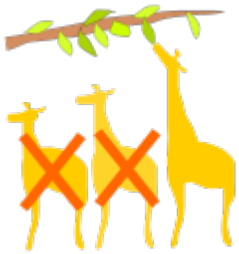
2. an **enumeration** corresponding to a list of at least of two co-items which may enter into various discourse relations;
3. a **closure** which closes the enumeration and may contain the enumeration's macro-Theme or an **encapsulation** *i.e.* a summarising anaphoric expression (*e.g.* Among [these three segments], Such factors ...).

While enumeration is a necessary component, trigger and closure are optional. Example (2) shows a prototypical enumerative structure extracted from the ANNODIS corpus (a Wikipedia(fr) article about Natural selection, emphasis is ours):

Figure 2: Capture of example (2)

Principe 2: Les individus les plus adaptés au milieu survivent et se reproduisent davantage [heading]

Certains individus portent des variations qui leur permettent de se reproduire davantage que les autres, dans un environnement précis. On dit qu'ils disposent d'un avantage sélectif sur leurs congénères:



- **La première possibilité** est, par exemple, qu'en échappant mieux aux prédateurs, en étant moins malades, en accédant plus facilement à la nourriture, ces individus atteignent plus facilement l'âge adulte, pour être apte à la reproduction. Ceux qui ont une meilleure capacité de survie pourront donc se reproduire davantage.

- **Dans le cas particulier de la reproduction sexuée**, les individus ayant survécu peuvent être porteurs d'un caractère particulièrement attirant pour les partenaires de sexe opposé. Ceux-là seront capables d'engendrer une plus grande descendance en copulant davantage.

Dans les deux cas, l'augmentation de la capacité à survivre et à se reproduire se traduit par une augmentation du taux de reproduction et donc par une descendance plus nombreuse, pour les individus porteurs de ces caractéristiques. On dit alors que ce trait de caractère donné offre un avantage sélectif, par rapport à d'autres. C'est dans *ce principe d'adaptation* uniquement, qu'intervient le milieu de vie.

This example, which covers an entire subsection, may be represented in its entirety by an enumerative structure. The heading and first paragraph introduce the Theme of *selective advantage* (“avantage sélectif”). The final punctuation of this first paragraph (“:”) signals that it opens out onto a new segment. It forms the trigger of the enumerative structure. The next two paragraphs, each

introduced by a bullet and an enumerative marker (*The first case*: “la première possibilité” and *In the particular case of sexual reproduction*: “dans le cas particulier de la reproduction sexuée”), are the two co-items which together compose the enumeration. Finally, the last paragraph closes the enumeration, encapsulating the co-enumerability criterion *via* the sentence-initial adverbial: *in both cases* (“dans les deux cas”).

Enumerating can be seen as a basic strategy in written expository text, the genre selected for the ANNODIS corpus, where writers want to express and readers except to read “unfolding ideas, claims, and arguments in terms of the logical interrelations among them” (Berman & Nir-Sagiv 2007:80). Of particular interest to our annotation project is the fact that enumerative structures occur at all levels of granularity and can be nested. Enumerative structures are capable of handling the global texture of a text as well as the very local texture at sentence level, as seen in example (3), which provides, with a larger extract from the Wikipedia article, the co-text for example(2).

Figure 3: Capture of example (3)

This section is entirely constructed around a global enumerative structure, signalled by a wide range of cues. The following enumeration lists these cues, from the higher level ones to the more local:

1. Headings “package” this enumerative structure through their hierarchical relationships, syntactic parallelism and semantic contents.
2. In parallel to this organisation *via* headings, pictures add a dimension of visual parallelism between subsections.
3. The first sentence ends with a typical trigger pattern: V PP where NP is composed of a numeral and plural N (*is based on three principles*: “repose sur trois principes”).
4. The first paragraph is a pre-enumeration which acts as the trigger for the main enumerative structure.
5. Each subsection ends with a presentational sentence focusing on an italicized NP which names the principle just explained (*e.g.* it is the *variation principle*)
6. The first sentence of the last subsection starts directly with a strong co-item cue: *The third condition for natural selection* (“la troisième condition pour qu'il y ait sélection naturelle)”). This clearly signals this subsection as third co-item of a list.
7. Finally, the last paragraph of this extract starts with a prototypical encapsulation: *These third principles* (“Ces trois premiers principes”), signalling the start of the closure. This cue indicates that the last paragraph has to be interpreted at the same level as the first paragraph *i.e.* at the highest level.

Let us stress that enumerative structures are not just a matter of layout or text formatting. In example (1), sentence-initial time adverbials may be considered as cues delimiting co-items and lexical reiterations of the pair debate/discussion emphasise the hyper-Theme (*i.e.* the co-enumerability criteria). Nevertheless, this interpretation is strongly helped by layout in cases where sentence-initial adverbials are in fact paragraph-initial adverbials. Example (4) illustrates a “softer”

form of signalling: the enumerative structure is not visually marked, but signalled by an association of three cues: structural parallelism - a string of discourse frames (higlighted) -, topical chaining (in bold), and an encapsulation summarising the first three sentences in one referential expression (These three directions (“Ces trois directions”)).

Example (4)

4.2. Le Liban [heading]

La situation intérieure libanaise connaît un blocage institutionnel et politique caractérisé par une forte polarisation entre deux blocs, s'appuyant chacun sur des alliés extérieurs.

Depuis septembre 2004, **la France** a pris la direction d'un mouvement diplomatique qui a conduit à l'adoption par le Conseil de sécurité de la résolution 1559 appelant au retrait des forces syriennes du Liban. Après l'assassinat de l'ancien Premier Ministre Rafic Hariri, **elle** a pris clairement position pour la coalition des forces politiques du 14 mars, un bloc dont le principal ciment et l'objectif commun étaient de mettre fin à l'influence syrienne au Liban. Au lendemain de la guerre d'Israël contre le Hezbollah à l'été 2006, **elle** a su mobiliser un large soutien international pour la mise en place d'une FINUL renforcée et pour la reconstruction du pays dévasté lors de la conférence de Paris 3. **Ces trois directions** , engagées au cours des trois dernières années, méritent un examen critique, au niveau des objectifs d'une part, du cadre dans lequel la France déploie son activité et des partenaires qu'elle choisit d'autre part, pour envisager les options politiques à venir.

These examples illustrate some of the ways in which enumerative structures may be used in expository text, and their ability to function at different scales. Our interest is clearly in the structuring principles underlying enumerative structures rather than in specific formatting devices, our objective being to use them as an “illustrative model” for the complex sequentiality phenomena we want annotated.

III. Enumerative structure as an annotation model

Enumerative structures are used in ANNODIS as a framework for representing discourse organisation and annotating continuity/discontinuity relations. Such a model based on enumerative structure strikes us as particularly well-suited to an annotation task thanks to its simplicity: it is intuitively familiar to naive coders. But its greatest asset for our objectives is that the identification of enumerative structures is guided by surface cues. Indeed, enumerative structures cannot be defined (and interpreted) independently of the cues that signal them. When encountering a trigger

pattern, the reader constructs an expectation of the discourse organisation ahead. Similarly, an expression such as The third point is... has an instructional meaning indicating that two points were developed previously, and that there is an enumeration in the process.

The importance attributed to surface cues is fundamental in our approach to discourse organisation, which gives a central role to the interaction between the perceptive properties of a document (layout, but also lexico-syntactic elements, their textual position, their environment, etc.) – its "form" -, and the semantic contents that readers interpret, its "meaning". We see in enumerative structures a particularly interesting object for the study of the articulation between the ideational and textual components, to use notions from Systemic Functional Linguistics(Halliday 1985).

IV. TECHNICAL REQUIREMENTS

V. Corpus requirements

The end purpose of the ANNODIS project is to build a corpus of written expository French texts. Added to the thorny problem of availability and copyrights, three main corpus-selection requirements follow on from our decision to view discourse organization from a top-down perspective in order to identify high-level structures. First, text genres must be carefully selected so as to include long expository texts, such as scientific papers or essays. Whereas it is possible for short texts to rely solely on thematic continuity, longer texts, and in particular non-narrative texts with no default structuring in terms of succession of events, require other forms of organisation and provide more interesting data for discourse annotation and analysis. Second, the corpus must be composed of texts in which crucial elements of discourse organization such as subdivisions and layout are available. As described by Power *et al.* (2003), there is a strong connection between textual layout and semantic content: the graphical overlay of the document is a key element to access the document structure. It is therefore necessary to keep track of the physical properties of the document, especially at the level of subsections and titled segments. Third, the corpus must present various sub-types in order to allow the description of a discourse phenomenon across

genres.

In view of these criteria, texts selected for inclusion in the corpus combine newspaper articles, encyclopaedia articles, scientific papers and reports in the field of geopolitics. The document is encoded in XML, following TEI-P5 encoding procedures.

VI. Marking procedure

In the ANNODIS project, human annotation is assisted by computational tools. In a stage we call "marking", corpus texts are automatically tagged for a wide range of cues associated with enumerative structures and discontinuity. Coders are guided by these marks, made highly visible in the texts with coloured highlighting, in the detection of enumerative structures. Some of these cues are directly linked to the components of an enumerative structure. Others may play a role in the signalling of discourse organisation in a less clear-cut manner: section headings, paragraph breaks, sentence-initial adverbials, connectors, instances of syntactic parallelism.

The marking procedure uses information obtained from automatic POS tagging (Treetagger), syntactic dependency analysis (Syntex – Bourigault 2007), and layout analysis (textual positions, *e.g.* first or last sentence of a paragraph, sentence following a heading, etc.).

Two types of procedure are carried out: a set of regular expressions detect discourse organisation cues and a module measures parallelism between chunks of text.

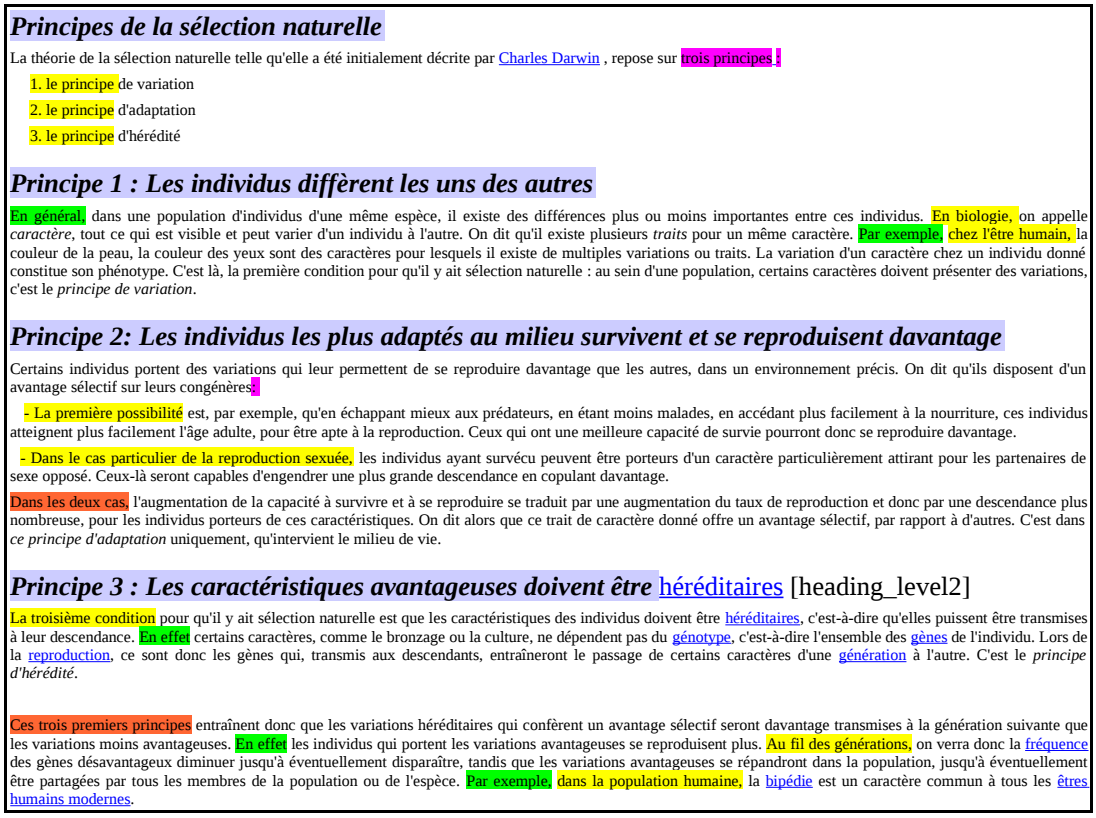
Four sets of cues are distinguished:

1. trigger cues: punctuational and lexico-syntactic patterns (*e.g.* a paragraph ending with [:], with an indefinite plural or a regular expression such as *as follow*), layout (*e.g.* headings);
2. co-item cues: punctuational and lexico-syntactic patterns (specific sentence-initial adverbials such as *first, ..., thirdly, finally* and also circumstantial sentence-initial adverbials or grammatical subjects such as *the second point ... the last characteristic* ...), layout (*e.g.* bullets, indentation, formatted lists) and syntactic parallelism;
3. closure cues: lexico-syntactic encapsulation patterns (*e.g.* demonstrative NP);
4. other cues relative to discourse organisation: sentence-initial connectors, co-referential

expressions in grammatical subject position (pronouns, reiterations, etc.).

Figure 4 shows example (3) with all cue marking according to this procedure (the pictures have been removed as in the ANNODIS annotation interface).

Figure 4: colour marking of example (3) in the ANNODIS annotation interface



Principes de la sélection naturelle

La théorie de la sélection naturelle telle qu'elle a été initialement décrite par Charles Darwin , repose sur trois principes :

1. le principe de variation
2. le principe d'adaptation
3. le principe d'hérédité

Principe 1 : Les individus diffèrent les uns des autres

En général, dans une population d'individus d'une même espèce, il existe des différences plus ou moins importantes entre ces individus. En biologie, on appelle *caractère*, tout ce qui est visible et peut varier d'un individu à l'autre. On dit qu'il existe plusieurs *traits* pour un même caractère. Par exemple, chez l'être humain, la couleur de la peau, la couleur des yeux sont des caractères pour lesquels il existe de multiples variations ou traits. La variation d'un caractère chez un individu donné constitue son phénotype. C'est là, la première condition pour qu'il y ait sélection naturelle : au sein d'une population, certains caractères doivent présenter des variations, c'est le *principe de variation*.

Principe 2: Les individus les plus adaptés au milieu survivent et se reproduisent davantage

Certains individus portent des variations qui leur permettent de se reproduire davantage que les autres, dans un environnement précis. On dit qu'ils disposent d'un avantage sélectif sur leurs congénères:

- La première possibilité est, par exemple, qu'en échappant mieux aux prédateurs, en étant moins malades, en accédant plus facilement à la nourriture, ces individus atteignent plus facilement l'âge adulte, pour être apte à la reproduction. Ceux qui ont une meilleure capacité de survie pourront donc se reproduire davantage.

- Dans le cas particulier de la reproduction sexuée, les individus ayant survécu peuvent être porteurs d'un caractère particulièrement attirant pour les partenaires de sexe opposé. Ceux-là seront capables d'engendrer une plus grande descendance en copulant davantage.

Dans les deux cas, l'augmentation de la capacité à survivre et à se reproduire se traduit par une augmentation du taux de reproduction et donc par une descendance plus nombreuse, pour les individus porteurs de ces caractéristiques. On dit alors que ce trait de caractère donné offre un avantage sélectif, par rapport à d'autres. C'est dans *ce principe d'adaptation* uniquement, qu'intervient le milieu de vie.

Principe 3 : Les caractéristiques avantageuses doivent être héréditaires [heading_level2]

La troisième condition pour qu'il y ait sélection naturelle est que les caractéristiques des individus doivent être héréditaires, c'est-à-dire qu'elles puissent être transmises à leur descendance. En effet certains caractères, comme le bronzage ou la culture, ne dépendent pas du génotype, c'est-à-dire l'ensemble des gènes de l'individu. Lors de la reproduction, ce sont donc les gènes qui, transmis aux descendants, entraîneront le passage de certains caractères d'une génération à l'autre. C'est le *principe d'hérédité*.

Ces trois premiers principes entraînent donc que les variations héréditaires qui confèrent un avantage sélectif seront davantage transmises à la génération suivante que les variations moins avantageuses. En effet les individus qui portent les variations avantageuses se reproduisent plus. Au fil des générations, on verra donc la fréquence des gènes désavantageux diminuer jusqu'à éventuellement disparaître, tandis que les variations avantageuses se répandront dans la population, jusqu'à éventuellement être partagées par tous les membres de la population ou de l'espèce. Par exemple, dans la population humaine, la bipédie est un caractère commun à tous les êtres humains modernes.

The visual appearance of these marks can be modified through the interface.

VII. Interface requirements

There are two major interface requirements which are crucial for the annotation process: one concerns document visualisation and the second one the annotation procedure itself.

VII.1.1. Document visualisation

The document visualisation interface must take into account text layout and formatting, and must allow highlighting of the cues marked automatically. An annotation model has been developed in order to record all this information in a uniform manner: each unit (heading, paragraph, pattern

trigger, sentence-initial adverbial, etc.) is associated with:

1. a feature set corresponding to the unit's properties, and
2. a positioning set corresponding to the absolute starting position and ending position of the unit in the text body.

In order to do this, the document has been split into two files: the layout and marking information goes in one file, the text body *i.e.* the succession of words and punctuation marks without any paragraph breaks in the other.

Once the information is transformed according to this process, the interface matches unit information with the text body and the delimitation of the macro discourse structures can be performed.

VII.1.2. Annotation of macro structures

Coders may have to annotate textual zones of any size, taking into account discontinuity and possible overlaps with previously delimited zones. As long as annotations are recorded according to the same model as layout and marking information, overlaps are not a concern.

We intend to perform inter-coder comparisons. In the XML file containing information on units, a set of meta data is automatically generated in order to record information relative to the coder (name) and the annotation date.

The interface offers a very simple way of delimiting and characterizing units with the help of a panel of editing tools. It also offers a palette of colours to characterise marked cues, as will be seen in the next section.

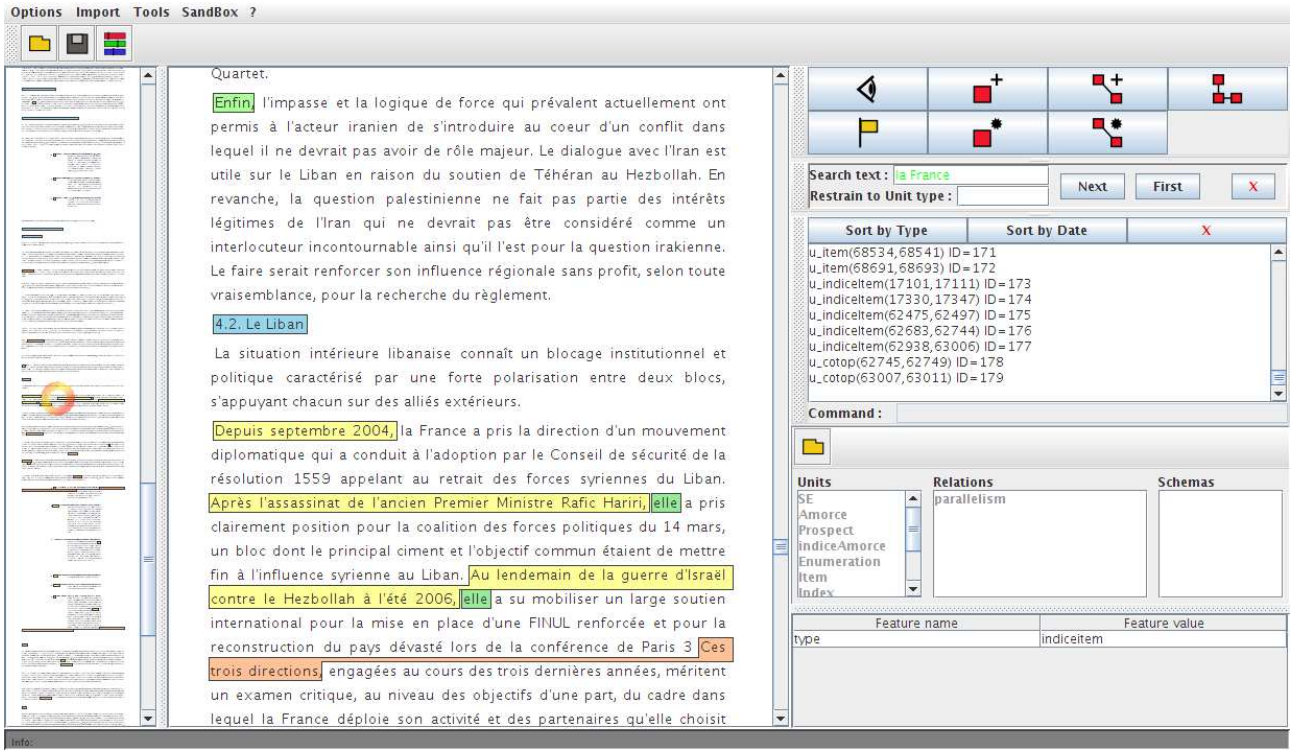
The last fundamental need the interface must satisfy concerns navigation inside the text, and more precisely the need to allow two concurrent points of view on the text. The interface allows a global view where configurations of marks and global layout appear, and a local view which enables coders to read the text and perform accurate annotations. Modes of navigation in these two views are of course coordinated.

VIII. THE ANNOTATION PROCESS

As described previously, coders are guided in their task by a wide range of cues highlighted with a specific colour, which allows them to distinguish classes of cue whether in a local or global view of the text. Moreover, an annotation guide is available where the different components of an enumerative structure are defined, illustrated and associated with the marked cues, with a number of basic tests to help detect them. All the annotation process is realised through the interface.

Figure 5 gives a screenshot of the ANNODIS interface specially developed for corpus annotation. On the left, a global view of the text appears in a “ribbon” with cues highlighted. The central window provides a readable view of the text where structures can be annotated. The little orange circle in the ribbon locates the text appearing in the readable view. Clicking on a zone in the ribbon brings the corresponding text in the readable view. The coder uses the ribbon to scan the text and identify dense zones of coloured cues, then clicks on a dense zone in order to bring it into view in the central window, and finally proceeds with delimiting and characterising the components of any enumerative structures present.

Figure 5: example (4) through the ANNODIS interface



The colour-marking should facilitate agreement in the definition (and recognition) of enumerative structures given that, as stated previously, enumerative structures essentially emerge from cues. Besides, without such colour-marking the coder would have no choice but to read the text entirely from beginning to end in order to detect enumerative structures; we fear that such a linear reading process would lead to large-scale highest-level structures (such as those covering several subsections) being missed out. Colour-marking allows another way of accessing text, starting from a global view which encourages a scanning/skimming approach and in due course zooming down to a more local view.

Through this dual way of accessing text, the interface embodies our hypothesis that readers blend a top-down and a bottom-up approach during reading. We guide them towards a top-down approach for scanning and towards a bottom-up approach for annotating.

IX. CONCLUSION

This paper ambitions to illustrate a macro approach to discourse organisation. To bring near this level of discourse, we have focused our attention on the strategies of continuity and discontinuity. More precisely, we have tried to point out that enumerative structures could be seen as meta structures covering a range of discourse organisation phenomena and providing a concrete illustration of a certain kind of sequentiality. Furthermore, because of their simplicity, enumerative structures are intuitively familiar to novice coders and could be identified by the way of surface cues that signal them. To attempt this aim, we have expounded the document visualisation interface required by the annotation process.

Finally, we hope that we have convincingly demonstrated that our annotation model is available for developing a diversified French corpus annotated with macro discourse structures. From now on, based on the plan of action presented here, we have to start the annotation stage in order to give some accurate and straight results of our approach.

X. REFERENCES

Berman R.A. & Nir-Sagiv B. (2007). Comparing Narrative and Expository text construction across adolescence: a developmental paradox. *Discourse processes*, 43:2, 79-120.

Bourigault D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*, Unpublished Habilitation Dissertation. Université de Toulouse 2, France.

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

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

Halliday M.A.K. (1985). *An introduction to Functional Grammar*. London: Edward Arnold.

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

Luc, C., Mojahid, M., Péry-Woodley, M.-P. & Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes, CIDE (Colloque International sur le Document électronique)*, Lyon, France, pp. 21-40.

Martin J.R. (2001). Cohesion and texture. In D Schiffrin, D Tannen & H Hamilton (Eds.) *Handbook of Discourse Analysis*. Blackwell: Oxford, pp. 35-53.

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

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