Discourse organisation through Theme position

Lydia-Mai Ho-Dac, Université Toulouse2 le Mirail, France

This paper focuses on the role of elements placed in the initial position *i.e.* elements fulfilling the role of Theme in discourse organisation. The large-scale corpus study proposes a new methodology based on quantitative analysis of the discourse roles of sentence-initial elements. The theoretically-based hypothesis is that Theme position has an important function in discourse organisation. Theme, defined as the starting point of the message, is composed of the first elements that the reader perceives. The analysis of the distribution and the use in discourse of these elements gives us a great overview on the textual organisation of different types of text.

Keywords: discourse organisation, Theme position, data-driven approach

1 Theme and discourse organisation

The object of this study concerns what I call 'discourse organisation' or 'text organisation'. These terms – considered here as synonymous – convey the idea that text is not a bag of words, a bag of sentences, a bag of paragraphs but should be seen as a structured object. This structure is the result of the strategies used by the reader when he produces his text. Two main strategies are identified in my study: strategy of continuity and strategy of discontinuity. The writer may want to link two units by a continuity relationship (strategy of continuity) or may want to indicate a shift between two units (strategy of discontinuity). One of the consequences of this is that text is segmented in interrelated 'chunks' which may be embedded in a hierarchical relationship.

My claim is that Theme plays a crucial role in signalling these two strategies. This claim is based on a positional definition of Theme that corresponds to the beginning of textual units (sentences, paragraphs, sections) and on the idea that the beginning of textual units is a good location to indicate if there is a discontinuity or a continuity. Moreover, during reading, we build some assumptions based on the first elements perceived, assumptions that may orient the interpretation of the rest of the unit. Theme position corresponds here to the entire preverbal zone as stated by Berry (1995), Fries (1995) or Mathiessen (1995). When I apply this definition on paragraphs or sections, Theme corresponds to the preverbal zone of the first sentence of the textual unit. This acceptation of Theme is necessary to take into account all the complexity of discourse organisation.

1.1 Discourse organisation, text segmentation and sequentiality

Discourse organisation and text segmentation are seen as the consequence of the 'linearization problem' (Levelt 1981). Although the representation we have in our mind is not linear (similar to a picture, a form, a scene, etc.), the text (either written or oral) must be linear. Text is a succession of sentences. Sentences must appear one after the other in time. One cannot write or speak, read or understand several sentences simultaneously. This lack of isomorphism between mental representation and what we must produce is at the root of discourse organisation. The study of discourse organisation aims to find answers to the question: how speaker and writer go about presenting information in a linear format.

I take as my starting point the issue of sequentiality in text as defined in Goutsos (1996) who sees text as a "*periodic alternation of transition and continuation spans*" (*op.cit.*:501). Model of Sequentiality distinguishes three levels of discourse structure. The cognitive level sees the writer's mental representation as structured by the basic strategies of continuity and discontinuity. The linguistic level is concerned with the techniques available to realize these strategies. The textual level is the material result of these strategies and techniques. Text segmentation into continuation and transition spans pertains to the textual level.

Text segmentation can be viewed from the continuity angle and the discontinuity angle. From the continuity angle, linguistic units cluster around a specific interpretation criterion. From the discontinuity angle, text is divided into segments or spans (in Goutsos' terminology). The criteria which bind text units together may concern parts of the subject matter of the text (e.g. referential continuity, time reference) or the presentation process (e.g. rhetorical or document structure). As long as a criterion remains valid, the segment is open and incoming linguistic units join into a 'continuation span'. When it is no longer valid, the segment is closed and the resulting discontinuity is signalled via a 'transition span' which indicates a shift in the discourse process. Such a shift may be, for example, a referential break, the end or opening of a discourse frame, a rhetorical articulation or the end or beginning of a document structure segment (paragraph or section). Continuity being the default, 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.

Example 1[1] constitutes a good illustration of a combination of continuation and transition spans. In this extract, a string of cohesion devices establishes continuity around the topic of "*debate between specialists of transatlantic relations*" (I put these devices 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 topical continuity.

(1) Since the end of the cold war, ***the debate between specialists of transatlantic relations*** *has tended to be satisfied with worthy pronouncements and much simplification.* ***It*** *has not shown sufficient concern for the breadth of the changes taking place [...].*
More recently, ***the discussion*** *has been focusing on a supposed gap in social values*

[1] The original French version is given in appendix.

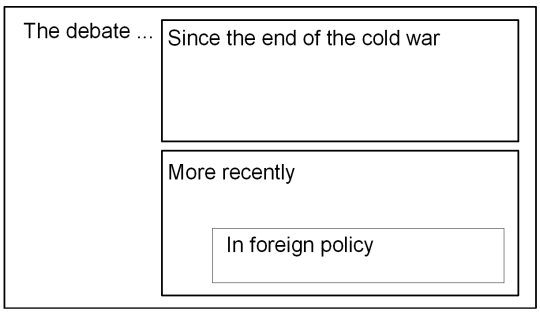
between the two shores of the Atlantic, an idea to which the events of 9/11 have put an end. ***This debate*** *is ongoing, but it is now limited to the domain of social analysis.* In foreign policy terms, ***this discussion on continental shift*** *has turned into an opposition between the unilateralism of America's policy and the multilateralism of its European partners.*

At the same time, example 1 shows three cues of shift opening new circumstance frames (these cues are not italicized). The first is a temporal frame introduced via an adverbial setting a time reference (*Since the end of the cold war*). This time reference remains valid all through the first paragraph. We can therefore say that all the sentences in this first paragraph cluster around a common time reference. The scope of this first time adverbial constructs a first (sub-)continuation span.

The second paragraph begins with another time adverbial *(More recently*) expressing a different time reference. This adverbial introduces a new temporal frame and then, signal a shift. As for the first frame, this one also spans the entire paragraph. We can also mention a third frame introduced with *In foreign policy terms.* This notional frame corresponds to the last sentence and fits inside the temporal frame introduced with *More recently* without closing the temporal frame.

Figure 1 gives a representation of this complex sequentiality. It shows that a continuation span may contain several other (sub-)continuation spans associated with a different component of the process (e.g. circumstances). In this example, transition spans are minimal, consisting in the space between two sentences or two paragraphs. To refer to these different components of the process, I use the Gestalt distinction between figure and ground. The participants or the topics of the process plays on figure whereas the circumstances of the process have to do with ground (by setting the scene in which the figure appears).

Figure 1: Representation of sequentiality in example 1



All the continuity cues (cohesion devices) and the discontinuity cues (initial adverbials) mentioned occur in Theme position. This positioning is not a coincidence. One of my work's basic claims is that elements in Theme position play a crucial function in signalling sequentiality in discourse *i.e.* in indicating whether there is discontinuity or continuity in the discourse flow.

1.2 The role of Theme in discourse organisation

Some cognitive studies (e.g. Enkvist 1989, Givón 1995) claim that what the writer expresses first corresponds to the 'crucial' information. 'Crucial information' means information necessary to the correct interpretation of the purpose of a given message. This concept is derived from the "*Crucial Information first*" principle (CIF) defined in Enkvist (1989). Enkvist opposes this principle to the "*Old Information First*" principle. Contrary to Enkvist, I do not oppose these two pragmatic principles. I think that crucial information can in some cases correspond to old information. It depends on the information flow. Either the writer wants to indicate that incoming information is in continuation with preceding information; or he wants to indicate that there is a shift or a break. In both cases, indication is given by the first elements of the message. In the case of continuation, first elements may correspond to given information. In the case of shift, first elements express information that orients the rest of the message by, for example, setting new circumstances (this conception is also supported by Berry 1995, Fries 1995 and Mathiessen 1995). From the point of view of experiential metafunction, elements in initial position express circumstances or entities implied in the process described. From the point of view of textual metafunction, first elements organize discourse by segmenting it into chunks *i.e.* by indicating if the segment is still opened or if it is closed.

According to my positional definition of Theme, I may say that Theme participates in the management of discourse organisation, by fulfilling a dual function: orientation and connection.

1.2.1 Theme as orientation

One discourse function of Theme consists in forward-looking orientation: Theme is the beginning of the message and, as a consequence, sets preliminary criteria of interpretation for the rest of the message. As stated by Fries:

> *"Theme functions as an orienter to the message. It orients the listener/reader to the message that is about to be perceived and provides a framework for the interpretation of that message*" (Fries 1995:318).

In terms of discourse comprehension, elements that have been read first have a stronger influence on the interpretation of the rest of the message than later elements (see Thompson 1985, Hasselgård 1996, Le Draoulec & Péry-Woodley 2003). If we focus particularly on initial elements occurring before the grammatical subject (such as initial adverbials), we find elements that set a discourse frame for the interpretation, as stated by Chafe:

> *"[elements in initial position] limit the domain of applicability of the main predication to a certain restricted domain [...] set[ting] the spatial, temporal or individual framework within which the predication holds"* (Chafe 1976:53)

This is typically the case with the three initial adverbials in the example 1. Chafe's notion of "*restricted domain*" is related to the notion of 'discourse frame'. Discourse

frames have emerged from several French studies initiated by Charolles (Charolles 1997, Charolles et al. 2005). These studies suggest that initial adverbials have an instructional meaning relative to segmentation by projecting an interpretation criterion forward and thus defining the initial boundary of a segment. In our view, two properties of adverbials are called upon in this definition: scope and structuring power. "Scope" corresponds to the semantic continuity of the reference expressed by the adverbial. "Structuring power" (cf. Le Draoulec & Péry-Woodley 2003, 2005, Ho-Dac & Péry-Woodley 2008) corresponds to the capacity of initial adverbials to divide information into blocks *i.e.* segments within which sentences cluster around an interpretation criterion, often but not necessarily the reference expressed by the adverbial.

Grammatical subjects may also function as orienters for the rest of the message. In example 1, the first subject sets the main topic of the entire paragraph (and even of the section) : the *discussion between* ... In this case, Theme may be related to the notion of aboutness or in Halliday's terms: "*that with which the clause is concerned*" (Halliday 1985:38).

1.2.2 Theme as connection

A second discourse function associated with Theme consists in backward-looking connection: Theme connects the rest of the message to the preceding discourse by expressing elements that allow the reader to integrate incoming information in a coherent way into the mental model in construction. As Halliday states:

> *"Theme is the peg on which the message is hung*" (Halliday, 1970:161)

Theme as a peg is seen as a cohesion device. The common strategy that a writer can employ to indicate that incoming information is linked to the preceding information is to express given information in Theme position, and new information in Rheme position. In example 1, all grammatical subjects (except the first) connect the sentence to the preceding one. In other words, they create topical continuity in this continuation span. Conversely, because continuity is the default, the absence of the initial adverbial could be interpreted as a cue for time reference continuity.

2 Configurations of cues for marking discourse organisation

In preceding sections, I have presented my theoretical point of view on the role of Theme in discourse organisation. My concern is more specifically to identify the cues that have the capacity to indicate continuity or discontinuity. My main interest is to discover the kind of cues that we must take into account in order to describe and understand the textual level of discourse organisation.

My claim is that discourse organisation is signalled by configurations of cues rather than by single markers. Moreover, writers and readers have to manage several levels of organisation which include thematic continuity but also time and space reference, rhetorical articulation and document structure. As discourse cues may simultaneously contribute to several of these interdependent levels, a global view of

discourse organisation is needed. A data-driven approach seems to be a relevant way to have such a global view.

The configurations of cues that I aim to discover can be defined as follows: the co-occurrence of element A with element B in a specific text position in a particular text-type leads the reader to interpret either continuity or shift between the preceding discourse segment and incoming information. In this definition, I mention three kinds of cues: the lexico-syntactic elements A and B, text position and text-type. The rest of this section describes these three categories of cues.

2.1 LEXICO-SYNTACTIC ELEMENTS

The set of lexico-syntactic elements taken into account corresponds to all the elements that occur in Theme position. I choose to define Theme position as the equivalent to the overall preverbal zone like Fries (1995), Mathiessen (1995) and Berry (1995) do. This choice derives from the global view of discourse organisation adopted in this study.

Theme must be considered as simultaneously orienting and connecting in order to take into account the complexity of sequentiality *i.e.* the fact that there are different levels of organisation (e.g. figure, ground or rhetorical structure). In example 1, elements occurring in the preverbal zone of the first sentence of the second paragraph (*More recently, the discussion*) participate in signalling that there is ground discontinuity (in time reference) and figure continuity (in topical reference). In example 2[2], the same phenomenon is observed: the section is organized around a long topical continuity marked by the co-referential grammatical subjects and around three temporal frames introduced by the time adverbials. Each time adverbial has a scope and marks a discontinuity between three homogeneous segments in terms of time reference.

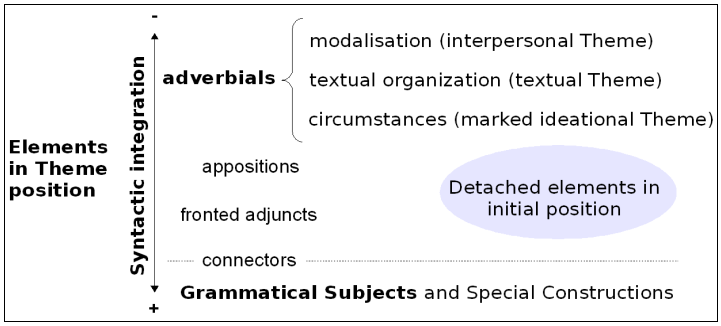
(2) ***Florence-Milan, 1500 - 1513*** *[heading]*
In 1500, ***Leonardo*** *goes to Mantova, where he draws Isabella d'Este's portrait, [...], to Venice, [...], and to Florence, where -[...]- he will stay till 1506.* ***He*** *shares his time between painting [...], and military engineering projects in the Arno valley and in Piombino.* ***Leonardo*** *resumes work on theTrattato started between 1487 and 1792, and continues until around 1513.* From 1506, ***he*** *divides his time between Milan where [...], and Florence where [...].* ***He*** *returns to his equestrian statue project, [...].* ***He*** *deploys an intense scientific activity: anatomy, mathematics, and produces architectural and decoration projects for Charles d'Amboise. But,* in 1513, ***he*** *leaves Milan for good as the city is reclaimed by the anti-French coalition.*

Rome-Amboise, 1513 – 1519 *[heading]*
In Rome, *where he has his lodgings in the Belvedere,* ***Leonardo*** *finds himself [...]*

Defining Theme position as the entire preverbal zone means the lexico-syntactic elements taken into account for the study are those represented in figure 2:

[2] The original French version is given in appendix.

Figure 2: Lexico-Syntactic elements taken into account



A first distinction is made between special and canonical constructions. Special constructions (or constructions with a 'special Theme') correspond to sentences where the grammatical subject has no referential meaning. These constructions may be used to focus on one part of the sentence (as in cleft constructions or left dislocations), to introduce new referents (presentational constructions) or to indirectly express an evaluation of the process being expressed (*it*-extraposition). In canonical constructions, the grammatical subject is the topical Theme.

A second distinction related to syntactic integration is made between detached elements and integrated elements. Detached elements are those which occur before the grammatical subject, separate or not with a comma (e.g. adverbials, appositions, fronted adjuncts). Integrated elements correspond to grammatical subjects.

2.2 Text-Type

Textual variation is a feature that must be taken into account in the description of sequentiality. It is one of the bases of my methodology: discourse organisation (and sequentiality in discourse) may be different in narrative, descriptive, argumentative, etc. texts. These different types of discourse organisation are related to the notion of text-type. This notion is defined here following Biber's view that types are determined by linguistic characteristics in opposition to genres which are identified with respect to extra-linguistic parameters such as social function (Biber 1988). From this perspective, there are strategies (as in Goutsos 1996) that are used more in certain types of text. These strategies depend on discourse organisation itself rather than on genre. For example, my corpus is composed with texts that have in common mono-referential property (they are constructed around a single topic). Conversely, there are pluri-referential texts strongly organized with space and time references. Section 3.1 describes my corpus and explains this text-type characterisation.

Text-type is a cue that works together with other more textual cues. In other words, the reader perceives (it is already interpretation) that the text corresponds to a particular type. On the basis of such a categorisation, he constructs assumptions relative to the construction of the text. Text-type has oriented his interpretation.

2.3 Text position

The last feature that forms the basis of my analysis is text position. Text position is linked to the level of organisation that is marked by document structure. Orientation and connection processes are likely to vary according to the level of organisation: whether the Theme starts a new section, a new paragraph, or merely a new sentence inside a paragraph.

This hypothesis is based on the idea that document structure strongly participates in constructing the meaning of a text. I consider that, from the reader's point of view, the beginning of a new section or a new paragraph triggers specific discourse processes that orient the interpretation. These discourse processes are not often studied in linguistics. In this study, I give a central role to document structure. Through 'playing' with the three text positions (in three text-types), I will discover configurations of cues relevant for discourse organisation.

The choice of three text positions follows a common (and quite intuitive) association between discontinuity and section or paragraph break. Therefore, the three text positions taken into account are:

- S1 = section initial position;
- P1 = paragraph initial position[3];
- P2 = paragraph internal position.

2.4 Some assumptions about these cues

According to the morpho-syntactic categories of grammatical subjects, the syntactic function of detached elements, and the different text positions, potential correlations can be derived between these cues and their contribution in indicating continuity or discontinuity in discourse organisation. The main assumptions are presented in table 1.

Table 1: Potential correlations between cues and their contribution in indicating discourse organisation

Correlation with	discontinuity	continuity
at experiential level		
figure	Grammatical subjects with a low degree of accessibility (Ariel 1990), paragraph breaks (P1), headings and new sections (S1)	Grammatical subjects with a high degree of accessibility (Ariel 1990), paragraph internal position (P2)
ground	circumstance adverbials,	apposition, fronted adjuncts,

[3] P1 is not taken into account when the paragraph begins a new section (S1 and P1 are exclusive).

Correlation with	discontinuity	continuity
	paragraph breaks (P1), headings and new sections (S1)	initial connectors
at rhetorical level	textual adverbials (or textual marked Theme), initial connectors, paragraph breaks (P1), headings and new sections (S1)	paragraph internal position (P2)

Grammatical subjects are represented here in terms of their degree of accessibility as in the scale devised by Ariel (1990). Accessibility models such as those of Ariel (1990), Gundel et al. (1993) or Centering Theory (Walker et al. 1998) aim to explain cognitive processes involved in the activation of discourse referents. One of these processes is concerned with the continued activation of a given referent. In Ariel's work, one way to keep a referent activated is to use morpho-syntactic elements that are correlated with a high degree of accessibility (*e.g.* pronouns, demonstrative NPs, etc.). Conversely, the introduction of a new referent must be accomplished via elements correlated with low accessibility (indefinite descriptions, new proper names, etc.). Ariel's accessibility scale enables me to classify referential expressions in a way that may correspond to the notions of continuity or discontinuity. The next section gives an overview of the proposed correspondences between degree of accessibility and grammatical subject morphology.

If we look back on example 1, we can see that all sentences have a topical Theme (*vs.* a special Theme). In the first sentence, the complete definite description correlates with a middle-low degree of accessibility (degree of accessibility = 2). This non co-referential expression sets the main topic of the section. In subsequent sentences, several cues of topical continuity occur: an anaphoric pronoun (with the highest degree of accessibility = 7), a reduced co-referential definite description (with a medium degree of accessibility = 3), a reduced co-referential demonstrative description (with a high degree of accessibility = 6) and finally a complete demonstrative description (with a high degree of accessibility = 5). Looking now at settings (*i.e.* ground), there are a number of discontinuity cues in the form of circumstance adverbials. These adverbials indicate to the reader that the frame has shifted: from one time reference to another, or into a specific domain of knowledge.

In example 2 extracted from texts of another text-type, anaphoric pronouns are seen to be more frequent. This frequent use indicates a strong topical continuity in this text. Time adverbials have the same function here as in example 1. These two examples are organized, for the figure, around a main continuation span and, for the ground, around three temporal frames. Strangely, the Theme of these two document structure segments (here sections) expresses a circumstantial reference and the topic of the segment. In example 1, each paragraph begins with a time adverbial and a referring expression related to the topic. In example 2, the first section (not divided into paragraphs) begins with a time adverbial and the second with a space adverbial. The

first grammatical subject of each section is the repeated proper name *Leonardo*. I suggest that these configurations of cues are meant to indicate the organisation of the section rather than a coincidence. It is this kind of configuration of cues that my methodology attempts to highlight.

3 A DATA-DRIVEN APPROACH

In accordance with this conception of discourse marking via configurations of cues, I set up a methodology that enables me to measure the contribution of the different kinds of cues: collocation of lexico-syntactic elements, text-type and text position. The basis of this quantitative analysis consists in experimenting the different kinds of cues. For each kind of cues, I measure the variation that its presence engenders in a data-driven approach.

> *"The problem with [a hypothesis-driven] approach is that during the investigation, we can search only for evidence, or lack of evidence, for what we expect to find. The alternative to hypothesis-driven research is data-driven research, in which we are informed by the corpus data itself and allow it to lead us in all sorts of directions, some of which we have never thought of."* (Rayson 2002:1)

The choice of a data-driven approach, necessarily based on an exhaustive analysis, aims to let the data 'do the talking' and to "trust the text" (cf. Sinclair 2004) contrary to a hypothesis-driven approach. As a result I chose to analyze all the elements in the preverbal zone rather than select elements for which there is a potential correlation (as presented in table 1). After describing my corpus, I briefly set out the automatic tagging that constitutes the starting point of several quantitative analyses before presenting my results.

3.1 CORPUS DESCRIPTION

The constitution of my (French language) corpus is determined by three choices. Firstly, I need to analyze long written texts because long written texts need a more complex discourse organisation than short texts or oral texts. Oral texts strike me as completely different as far as construction and interpretation are concerned. It is possible for short texts to work around a single topical continuity or around the default continuity established by human interpretation. For example, in texts under 2 pages, headings and section divisions are not need. The texts in my corpus are always over 10 pages in length and divided into sections.

The second choice follows from the first: I select expository texts where organisation is topic centered rather than narrative texts where organisation is participant and event centered. In expository texts, there is no relation of succession (as happens by default in narratives) or action structure that causes implicit organisation. Moreover, the use of headings is very rare and specific in narratives.

The third choice concerns the feature of textual variation. My corpus is composed with three sub-corpora representing three text-types relative to different subject-matter and presentational organisation.

– ATLAS (~205,000 words), composed of 3 descriptive social geography texts;
– GEOPO (~250,000 words), a collection of 32 argumentative texts in the domain of international relations;
– PEOPL (~220,000 words), 30 descriptive biographies.

Texts in ATLAS are much longer than in GEOPO and PEOPL. They are mostly organized in terms of space and time references acting as settings for large spans of text, with no strong topical continuity. Conversely, texts in PEOPL are organized around a strong topical continuity (the topic being the subject of the biography). All texts include parts structured around time, but temporal organisation is not the norm and never extends to the whole text. GEOPO is more difficult to characterize, with an occasional temporal organisation and rather weak topical continuities. If we count the frequency of nouns in each text, we see that in GEOPO and ATLAS, there is a wider variety of frequent nouns than in PEOPL. This difference could be interpreted as a cue to pluri-referentiality (many frequent nouns) and mono-referentiality (few frequent nouns). Concerning spatial reference, we see that basic space adverbials[4] (e.g. *In Europe*) are much more frequent in ATLAS than in GEOPO or PEOPL Moreover, we see that these space adverbials occur more often in initial position than elsewhere. Concerning temporal reference, ATLAS and PEOPL both display a high frequency of basic time adverbials in initial position (e.g. *In 1900*), much more so than in GEOPO. These few data constitute good support for my typology.

3.2 Automatic cue tagging

In order to perform an exhaustive analysis without selecting specific cues, and in accordance with my claim that initial position is an indicator of discourse organisation, I tagged all elements appearing in Theme position in every sentences in the corpus. This tagging is performed automatically for all the elements carried out in figure 2.

For each sentence (23.000 sentences), the program records the following features:
– its text position,
– its sub-corpus,
– the presence of a short detached connective (e.g. *But,* ...),
– the presence of one or more detached elements, and
– if the syntactic construction is canonical or special.

Each detached element (7022 numbered) is characterized in terms of:
– part of speech,
– function (circumstance adverbial, textual adverbial, apposition, etc.), and
– semantic meaning for circumstance adverbials (temporal, spatial, notional).

Finally, I proceed to the characterisation of grammatical subjects. Three properties are taken into account:
– its part of speech,
– its length (a distinction is made between short NP composing of less than four words and long NP composing of more than three words) and

[4] As this analysis is mostly based on automatic tagging, a basic expression must also be an expression which lends itself to automatic extraction.

– the fact that the NP's head repeats a noun already mentioned in the current section.

Figure 3: Scale of seven degrees of accessibility, from high (degree 7) to low (degree 0), adapted from Ariel's Theory of Accessibility (1990)

pronoun or possessive NP > demonstrative NP with lexical reiteration or short > long demonstrative NP without lexical reiteration > reiteration of a proper name ("redenomination") > definite NP with lexical reiteration or short > long definite NP without lexical reiteration > proper name without lexical reiteration > indefinite NP or special construction

I also adapt the accessibility scale of Ariel to my data as indicated in figure 3.

3.3 Measurement of variations

The data are systematically explored to search for configurations of cues *via* two main measures: deviations in the use of different linguistic elements in initial position, and degree of association between two elements occurring in initial position, detached elements and grammatical subjects.

The first step of the analysis consists in extracting lexico-syntactic cues which vary according to text-type and text position. To do that, I measure variations for each lexico-syntactic element between:

– distributions in each corpus compared with overall distributions;
– distributions in each text position compared with overall distributions.

To measure the significance of variations, I use the z-score /z/ that measures the distance between the raw score and the standard deviation. /z/ is negative when the raw score is below the mean, positive when above. I consider that a significant deviation is above or below 2.5. A deviation that corresponds to a z-score of +2,5 means that there is a 1% probability that this positive variation is attributed to chance.

The second step consists in measuring:

– variations in subject position which can be associated with the presence of a particular detached element;
– variations in detached position which can be associated with the presence of a particular type of grammatical subject.

These variations are measured in the host sentence and in the following sentence and by taking into account each text-type and each text position.

4 Results and interpretation

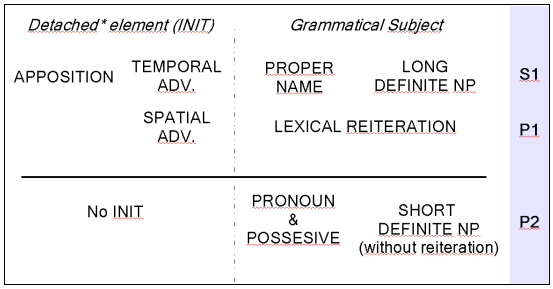
Through this exploratory method, I obtained a number of results that are presented in Ho-Dac (2007). Confronted with a multiplicity of data, it has not been easy to find a way to interpret and to represent these variations. In the first subsection, I present an

overview of the results obtained and their interpretation. I then go on to give an illustration of the method with a step by step account of the study of variations concerning time and space adverbials.

4.1 Organisation and text-types

The first set of results presented in figure 4 indicates the general associations showing a significant deviation according to text position. Figure 4 displays all the elements occurring in the preverbal zone for which the z-score test shows a significant association (/z/ > +2,5) with S1, P1 or P2. The label of all the elements that occur significantly more in section-initial or paragraph-initial and significantly less in paragraph-internal position is indicated above the horizontal line. Conversely, below the line are indicated all the elements occurring significantly more in paragraph-internal position and less in section and paragraph initial position.

Figure 4: Significant general associations between lexico-syntactic elements and text-position



If we focus on grammatical subjects, we find well-known associations. Categories that may strongly mark continuity such as pronouns and possessives occur significantly more in paragraph internal position (P2). On the other side of the horizontal line, there are elements traditionally linked to discontinuity such as:

- in S1, full definite descriptions and new proper names that may mark discontinuity by introducing a low accessible referent;
- in P1, lexical reiteration that may be used to emphasise a topical continuity when there is a shift in ground information or in rhetorical structure.

No significant variations according to text position are measured for special constructions. It seems that special constructions play a role in information structure rather than in global organisation.

For detached elements, there are associations between (i) absence of detached elements and paragraph-internal position (P2), and (ii) presence of detached elements and the beginning of document structure segments (S1 and P1). Appositions and time adverbials are significantly more associated to S1 in all corpora (despite a weaker /z/ in PEOPL concerning apposition). In P1, there are significant variations according to text-type: paragraphs seem to be organized around space references in ATLAS and around time references in GEOPO. In PEOPL, appositions, that serve topical

continuity, occur significantly more in P1. Only the strongest deviation, concerning space adverbials in ATLAS, is reported with general variations indicated in figure 4.

Table 2 summarizes the different significant variations measured for detached elements according to text-position in each sub-corpus.

Table 2: Detached elements: significant variations according to text position in each text-type

	S1 ----------> P1	P1 ----------> P2	
GEOPO	apposition time adv.	time adv.	no marked Theme
ATLAS		circumstantial adv. space adv.	
PEOPL	time adv. (apposition)	apposition	

The same measures for grammatical subjects are given by table 3.

Table 3: Grammatical subjects: significant variations according to text position in each text-type

	S1 ----------> P1	P1 ----------> P2	
GEOPO	definite NP	long (definite) NP	pronoun possessive NP short NP
ATLAS	definite NP long NP	lexical reiteration	
PEOPL	definite NP long NP proper name	repeated proper name	repeated proper name pronoun possessive NP

Variations measured for grammatical subjects may be interpreted with respect to the management of referential continuities in these three text-types. Whereas continuity seems to be achieved with lexical reiteration in ATLAS, GEOPO relies on reduced description. In PEOPL, the majority of proper names and pronouns signal strong topical continuity around a single topic (the famous person whose life story the text tells). Repeated proper names are associated here with high accessibility despite the fact that they are located in the middle of the accessibility scale. In fact, the status of repeated proper names is very characteristic in PEOPL. As Schnedecker (2005) showed, repeated proper names in biographies function more as alternatives to pronouns than as shift markers. This hypothesis is effectively supported by the significant association with P2.

Without going deeper into the data analysis (which is fully presented in Ho-Dac 2007) I propose to give an overview of the processes involved in this data-driven approach by focusing on space and time adverbials. Studying these two lexico-syntactic

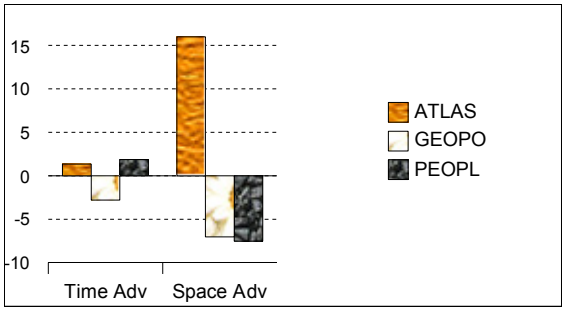
elements has enabled me to illustrate each step of my exploratory study of discourse organisation.

4.2 An illustration: variations associated with time and space adverbials

4.2.1 Step 1: variations according to text-type and text position

Time adverbials are frequent in detached initial position. They constitute 21% of all initial elements in our corpus. Space adverbials are less frequent (7% of all initial elements). All 1466 time adverbials and 500 space adverbials were analyzed. 31% of time adverbials are found in ATLAS, 36% in GEOPO and 34% in PEOPL; 66% of space adverbials are found in ATLAS, 21% in GEOPO and 13% in PEOPL. The results of the z-score test used to compare the distribution of elements in each text-type *vs.* overall distribution are given in figure 5.

Figure 5: Time and space adverbials: deviations acc. to text-type

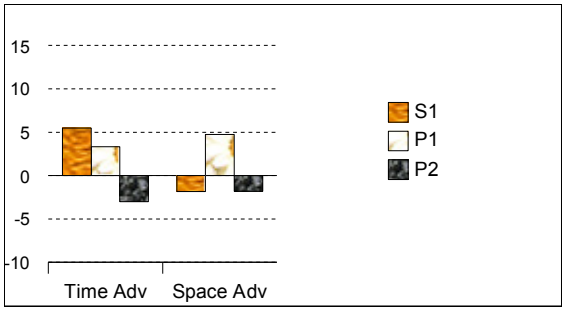


This figure indicates that time adverbials are not specific to one corpus. If they were specific to one text-type, there would be a positive significant deviation in one or more sub-corpus. In GEOPO, which is the least specific sub-corpus, there is a weakly significant negative /z/. This lower incidence means that there is a wider variety of initial elements in GEOPO rather than fewer time adverbials. In fact GEOPO has the highest number of occurrences of time adverbials : 522 compare to 452 in ATLAS and 492 in PEOPL.

Conversely, space adverbials seem to characterize ATLAS. The strong positive deviation associated with the two significant negative deviations means that the majority of space adverbials come from ATLAS.

Variations concerning text positions (the text positions are S1 : section-initial, P1 : paragraph-initial, P2: paragraph-internal) are given in figure 6. Here, the z-score test compares the distribution of elements in each text position *vs.* overall distribution.

Figure 6: Time and space adverbials: deviations acc. to text position



Time adverbials can be seen to occur significantly more in the first sentence of a section and in the first sentence of a paragraph while space adverbials occur significantly more in P1 only. Conversely, there are significantly fewer time adverbials in intraparagraphic sentences. The difference between space and time adverbials may be explained in terms of local discourse function *vs.* global discourse function. Space adverbials are associated with paragraph initial position but not with section initial position. Moreover, the deviation as regards space adverbials in P2 is not significant. This means that it is not rare to find a space adverbial in paragraph internal position, in contrast with time adverbials. These results may indicate that space adverbials fulfil a more local discourse function than time adverbials. The next results will confirm these observations.

Figure 7 indicates the results of the same z-score test applied in each sub-corpus in order to know if these associations with text positions are stable across the three text-types.

Figure 7: Time and space adverbials: deviations acc. to text position in each sub-corpora

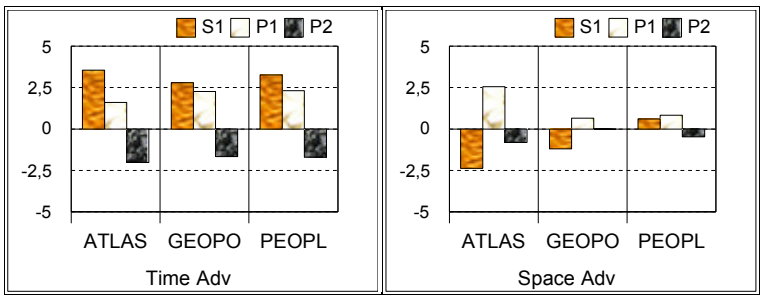


Figure 7 shows that all three corpora display the deviations associated with time adverbials: more time adverbials in S1 and P1 and fewer in P2. The z-score in P1 for GEOPO and PEOPL may be accepted as significant deviation, but not in ATLAS. Deviations concerning space adverbials are only linked to ATLAS in comparison with time adverbials. The discourse function of circumstance adverbials seems to be more definite in this sub-corpus. Time adverbials begin sections while space adverbials begin

paragraphs. This role distribution is facilitated by the fact that, in ATLAS, sections are very short and hierarchically embedded compared with PEOPL or GEOPO.

These first results lead us to conclude that adverbials may constitute good discontinuity markers because of their strong association with the starting point of document structure segments. I must however be careful in my interpretation of data. Firstly, this association has to take into account text-types (time adverbials appear to be less specific than space adverbials). Secondly, this association does not mean that adverbials signal discontinuity on their own for example, when they appear in non section or paragraph initial position. To clarify this last point, let us observe the lexico-syntactic environment of adverbials.

4.2.2 Step 2: variations in grammatical subject preceded by an adverbial

This analysis corresponds to the second stage of my methodology. I measured variations in subject position which can be associated with the presence of an adverbial in detached position. All the results are given in Ho-Dac & Péry-Woodley (2008). In this paper, I will only present the interpretation of these results.

Firstly, the discourse function of space/time adverbials seems to be very sensitive to text position. It seems that space/time adverbials are good segmentation markers when they occur in S1 or P1. In P2, it seems that the discourse function of space/time adverbials depends of the textual strategy used. Text may be organized around a dominant topical continuity or a dominant space/time structure.

Variations measured for grammatical subjects according to text-type and text position indicate that PEOPL is organized around a strong topical continuity unlike ATLAS and GEOPO, as was seen in the previous sub-section. The power of this topical continuity is also relevant in variations observed in host and following sentences of space/time adverbials and may explain the strong difference which opposes ATLAS and GEOPO to PEOPL.

In ATLAS and GEOPO, space/time adverbials may indicate discontinuity but only in specific configurations. Space/time adverbials collocate significantly more with reiterations that correlate with medium accessibility. This kind of subject may be used to emphasise a topical continuity when there is a shift in the setting (*i.e.* ground) or the rhetorical structure but not in thematic structure (*i.e.* figure). Space/time adverbials also collocate significantly more with new proper names that correlate with lower accessibility. This collocation may indicate that there is simultaneously a ground and a figure discontinuity. But variations in the following sentences do not support this suggestion. If the opening of a new time or space continuation span corresponds to the opening of a new thematic continuation span, the subject in the following sentence should correlate with high accessibility. But my data shows that it is not the case.

Grammatical subjects of sentences that follow a P2 sentence introduced with a space/time adverbial are significantly more associated with the bottom or the middle of the accessibility scale. We can also notice a significant association with demonstrative NPs. Demonstrative NPs correlate with high accessibility, but they mean more than just referential accessibility. The preferential use of demonstrative NPs in comparison to the use of pronouns is often associated, in French, with "reclassification". Reclassification consists in expressing a known referent stripped of its initial circumstances (De Mulder 1997). The referent's reclassification negates the possibility

of an extension of the adverbial's scope (see Ho-Dac & Péry-Woodley 2008 for more details).

The significant positive variations observed in the sentence following a space/time adverbial's host sentence may be a sign that space/time adverbials do not open a new continuation span at ground level. They merely locate the process of the host-sentence. Configurations where the host sentence's subject is a new proper name and the following sentence's subject is a demonstrative NP may indicate a discontinuity to do with the figure but not with the ground.

The case of PEOPL is very different. The topical continuity is so strong in this text-type that time adverbials seem to align their behaviour with the organisation established by topical continuity. In PEOPL, time adverbials co-occur significantly more with high accessibility co-referential expressions such as pronouns, possessive NPs and repeated proper names.

These associations in PEOPL are in agreement with the general model: in P2 subject referents present a remarkably high degree of accessibility, indicating topical continuity. This continuity is absolutely not disturbed by the presence of a time adverbial in initial position. The power of topical continuity is so strong that it is possible to have such associations in section initial or paragraph initial positions. This result agrees with observations presented in Le Draoulec & Péry-Woodley (2003) where authors show that, in narrative texts, time adverbials do not open a discourse frame but rather locate the chronological starting point for a succession of events. This is exactly what we have with the first time adverbial in example 2 where *In 1500* does not really extend its semantic scope until the second time adverbial. The semantic criterion of the first temporal frame is *from 1500 to 1506* instead.

Nevertheless, we may state that in example 2, time adverbials structure the text by indicating the boundaries of the three periods of Leonardo's life between 1500 and 1513. But this structuring power would certainly be less strong without this heading and if the section did not begin with a time adverbial predicting a time organisation for the rest of the document structure segment.

5 Conclusion

There are two aspects to my conclusion: the first concerns methodology while the second concerns advances in the study of discourse organisation.

Concerning methodology, I have shown the capacity of the data-driven approach to provide new insights. Firstly, it has proved to be a good tool for manipulating data. The z-score test is very simple to manipulate and enables us to test the structuring power of each feature that may interact in the signalling of discourse organisation. Secondly, it offers new perspectives for the study of discourse organisation. It enables us to identify the textual characteristics of global organisation. For example, ATLAS seems to have a strong spatio-temporal organisation while PEOPL has a more topical organisation. Thirdly, this methodology makes it possible to test the discourse function of specific lexico-syntactic elements such as time adverbials. Finally, this kind of study is very motivating for further discourse studies. As we have seen, there are other elements that are associated with document structure segments. It would be interesting

to describe their discourse function in a same way as we have just done for space/time adverbials.

Concerning now the advances in the study of discourse organisation, two points are essential. First of all, the hypothesis concerning the marking of discourse organisation has been partially validated. It is clear that we cannot speak about the structuring power of a lexical marker by itself. It is rather a matter of configurations of cues where lexico-syntactic elements play a role. These configurations seem to be quite complex. The case of time adverbials is a good illustration of this complexity.

This validation shows also that discourse organisation is strongly sensitive to text-type. In my study, considering text-type corresponds to taking into account the shape of a document and the textual strategies used in the document. A promising future direction would be to test the use of the configurations of cues discovered in this study in automatic text-type profiling.

References

Ariel, Mira. 1990. *Accessing noun phrase antecedents*. London: Routledge.

Berry Margaret. 1995: Thematic options and success in writing. In M. Ghadessy (ed) *ThematicDevelopment in English Texts* 55-84. London: Pinter, pp. 55-84.

Biber, Douglas. 1998. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Chafe, Wallace L. 1976. Giveness Contrastiveness definiteness subjects topics and point of view. In C.N.Li (ed) *Subject and Topic* 25-55. New-York: New York: Academic Press.

Charolles, Michel. 1997. L'encadrement du discours; univers champs domaines et espaces. *Cahier de Recherche Linguistique* 6. Nancy: LanDisCo université Nancy2.

Charolles, Michel, Le Draoulec Anne, Péry-Woodley Marie-Paule & Sarda Laure. 2005. Temporal and spatial dimensions of discourse organisation. *Journal of French Language Studies* 15(2). 115-130.

De Mulder, Walter. 1997. Les démonstratifs : des indices de changement de contexte. In N.Flaux, D.Van de Velde & W.de Mulder (eds) *Entre général et particulier : les Déterminants*. 137-200. Besançaon: Artois Press Université.

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

Fries, Peter H. 1995. Themes, Methods of Development, and Texts. In R.Hasan & P.H. Fries (eds) *On Subject and Theme. A Discourse Functional Perspective*. 317-360. Amsterdam: John Benjamins.

Givón, Talmy. 1995. *Functionalism and Grammar*. Amsterdam/Philadelphia: John Benjamins.

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

Gundel, Jeannette K., Hedberg N. & Zacharski R. 1993 Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274-307.

Halliday, Michael A.K. 1985. *An introduction to Functional Grammar*. London : Edward Arnold.

Halliday, Michael A.K. 1970. Language structure and language function. In R.Hasan & P.Fries (eds) *New horizons in Linguistics*. 140-164. Harmondsworth: Penguin.

Hasselgård Hilde. 1996. *Where and When: Positional and functional conventions for sequences of time and space adverbials in present-day English*. Oslo: Scandinavian University Press, Acta Humaniora.

Ho-Dac, Lydia-Mai. 2007. *Exploration en corpus de la position initiale dans l'organisation du discours*. Thèse de doctorat en sciences du langage. Université de Toulouse 2.

Ho-Dac, Lydia-Mai & Péry-Woodley Marie-Paule. 2008. Temporal adverbials and discourse segmentation revisited. *Linearisation and Segmentation in Discourse. Multidisciplinary Approaches to Discourse 2008 (MAD 08)*, Feb 20-23 2008, Lysebu, Oslo.

Le Draoulec, Anne & Péry-Woodley Marie-Paule. 2003. Time travel in text: Temporal framing in narratives and non-narratives. In L.Lagerwerf, W.Spooren & L.Degand (eds) *Determination of Information and Tenor in Texts* : *MAD 2003*, 267-275. Amsterdam & Münster: Stichting Neerlandistiek & Nodus Publikationen.

Le Draoulec, Anne & Péry-Woodley Marie-Paule. 2005. Encadrement temporel et relations de discours. *Langue Française* 148. 45-60.

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

Matthiessen Christian. 1995. THEME as an enabling resource in ideational 'knowledge' construction. In M.Ghadessy (ed) *Thematic Development in English Texts*. 20-54. London: Pinter

Rayson Paul E. 2002. *Matrix : a statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. in Computer Science, Lancaster University, UK.

Schnedecker, Catherine. 2005. Les chaînes de référence dans les portraits journalistiques: éléments de description. *Travaux de Linguistique* 5. 85-133.

Sinclair John McH. 2004. *Trust the Text: Language Corpus and Discourse*. London: Routledge.

Thompson Sandra. 1985. Grammar and Written Discourse: Initial vs. Final Purpose Clauses in English. *Text* 5. 55-84.

Walker, Marilyn A., Joshi Aravind & Prince Ellen. 1998. Centering in Naturally Occuring Discourse:An Overview. In M.Walker, A.Joshi & E.Prince (eds) *Centering Theory of Discourse* 1-28. Oxford: Calendron Press.

Appendix

Original French version of example 1.

(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 [...]*
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.*

Original French version of example 2.

(2) ***Florence-Milan, 1500 - 1513*** *[heading]*

En 1500, ***Léonard*** *se rend à Mantoue, où il dessine le portrait d'Isabelle d'Este, [...], à Venise, [...], et à Florence, où - [...]- il va rester jusqu'en 1506.* ***Son activité*** *se partage entre des travaux de peinture : [...], et des travaux d'ingénieur militaire dans le val d'Arno et à Piombino.* ***Léonard*** *remet en chantier le Trattato commencé entre 1487 et 1492, et y travaille jusque vers 1513.* À partir de 1506, ***il*** *partage son temps entre Milan où [...], et Florence, où [...].* ***Il*** *revient au projet de statue équestre, [...].* ***Il*** *déploie une grande activité scientifique : anatomie, mathématique, et fournit des projets d'architecture, de décors pour Charles d'Amboise. Mais,* en 1513, ***il*** *quitte définitivement Milan reconquis par la coalition antifrançaise.*

Rome-Amboise, 1513 – 1519 *[heading]*
À Rome, *où il loge au Belvédère,* ***Léonard*** *se trouve [...].*