An exploratory data-driven analysis for describing discourse organization

Lydia-Mai Ho-Dac

An exploratory data-driven analysis for describing discourse organization

Lydia-Mai Hồ-Đắc
Université de Toulouse-UTM

1. Introduction

Discourse or text organization is usually difficult to study with corpus linguistics methods because of the apparent incompatibility between the qualitative nature of discourse analysis and the quantitative requirements of corpus linguistics:

> "[I]t has been much less common to study discourse organization from a corpus perspective. In fact, these two subfields have research goals and methods that might be considered incompatible: The study of discourse organization – linguistic structure 'beyond the sentence' – is usually based on detailed analysis of a single text, resulting in qualitative linguistic description of the textual organization. In contrast, corpus studies are based on analysis of all texts in a corpus, utilizing quantitative analysis measures to identify typical distributional patterns that occur across texts." (Biber et al. 2007: 2)

The methodology we present here on a French corpus provides an alternative solution to this incompatibility, allowing a data-driven approach to discourse organization. It aims at describing high-level discourse structures (beyond the sentence) and discourse markers via a quantitative analysis based on automatic and systematic marking of linguistic features in a fairly large corpus of long written texts (c. 700.000 words in total). Three basic principles govern the methodology: it applies to long expository texts where discourse organization is necessary; it is based on the premise that discourse signalling results from an interaction between several forces and may also concern extra-linguistic features; finally, only elements occurring in sentence-initial, paragraph-initial and section-initial position are taken into account.

Section 2 situates our research question, which is the signalling of discourse organization and the discourse function of elements in initial position. Section 3 then exposes the range of cues taken into account. Section 4 details each step of the methodology that we need in order to describe the data configurations in section 5.

2. Discourse organization viewed through initial position

2.1. Discourse organization, text segmentation and sequentiality

The object of this study concerns 'discourse organization' or 'text organization.' 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. Discourse organization is 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 because sentences must appear one after the other in time. This lack of isomorphism between mental representations and what we must produce or what we have to interpret constitutes a major problematic in the study of discourse organization.

The issue of sequentiality in text as defined in Goutsos (1996: 501) proposes a solution by seeing text as a "periodic alternation of transition and continuation spans." His 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). Although Goutsos considers only thematic (dis)continuity, we argue that the specific interpretation criteria which bind text units together may concern different levels of organization: concerning part of the subject matter (e.g. thematic continuity but also space/time reference) or the presentation process (e.g. rhetorical or document structure). A shift between two segments may be a referential break, the end or opening of a discourse frame, a rhetorical articulation or the end or beginning of a paragraph or section.

Example 1[1] constitutes a good illustration of multi-level discourse sequentiality. In this extract, a string of cohesive devices establishes continuity around the topic of "debate between specialists of transatlantic relations." All these devices (in italics) may be interpreted as 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] We have translated examples from our French-language corpus to help comprehension.

(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 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 has three shift cues opening new circumstance frames (in bold type). First a temporal frame is 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. This scope effect builds, at the textual level, a first (sub-)continuation span.

The second paragraph begins with another time adverbial *(More recently*) expressing another time reference. This adverbial introduces a new temporal frame and signals a shift. As previously, this frame covers the entire paragraph. There is also a third frame – a "notional" frame this time –, introduced with the adverbial *In foreign policy terms*, fitting inside the temporal frame introduced with *More recently*.

Figure 1 gives a representation of this complex sequentiality: a continuation span containing several other (sub-)continuation spans associated with a different component of the process (e.g. topics and circumstances). The Gestalt figure/ground distinction helps us define the different components of the process. The participants or the topics of the process are seen as "figure" whereas the circumstances of the process have to do with "ground" (by setting the scene in which the figure appears).

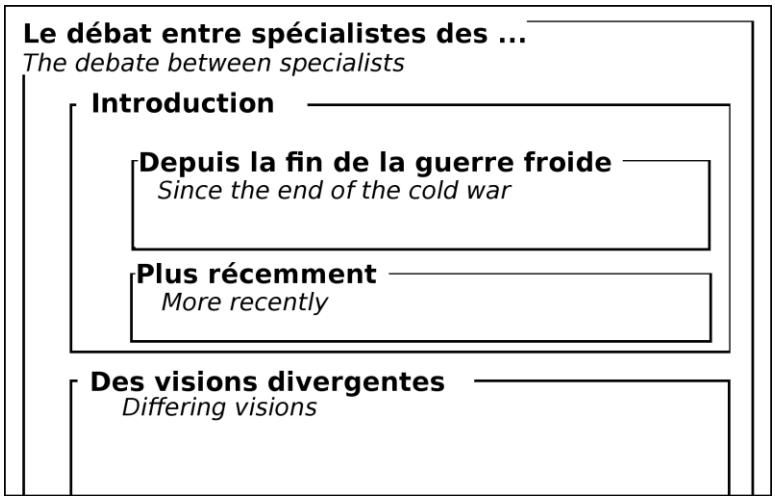

Figure 1. Representation of sequentiality in example 1

In this example, transition spans are minimal, consisting in the space between two sentences or two paragraphs. All the continuity cues (cohesive devices) and the discontinuity cues (initial adverbials) mentioned occur in sentence-initial position. This positioning is not a coincidence but a correlation between initial position and the signalling of sequentiality in discourse—that is, the indication of whether there is discontinuity or continuity.

2.2. The role of elements in initial position in discourse organization

Some cognitive studies (e.g. Enkvist 1989; Givón 1995) claim that what the writer expresses first corresponds to 'crucial' information, i.e. 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 by Enkvist (1989) as a counterpart of the *old information first* principle. According to information flow, crucial information can correspond to old information or to new information: 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, this indication is given by the elements coming first in the message. In the case of continuation, these first elements may correspond to given information. In the case of shift, the first elements express information which provides an orientation with respect to what comes next, for example, setting new circumstances. In other words, elements in initial position participate in the management of discourse organization by fulfilling a dual function: orientation and connection. These two functions define what is called *theme* in Systemic Functional Linguistics (Halliday 1985).

The first function associated with elements in initial position consists in a backward-looking connection: elements in initial position connect 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 (1970: 161) states: "Theme is the peg on which the message is hung."

In this way, elements in initial position are cohesive devices. For example, a common strategy used 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 (this is the old information first principle). In example 1, all grammatical subjects (except the first) connect the sentence to the preceding one, creating a topical continuity in this continuation span. Sentence-initial time adverbials also connect the first paragraph to the second paragraph by establishing a temporal organization between them.

At the same time as they establish a connection, elements in initial position orient discourse by setting preliminary interpretation criteria for the incoming message. Because they have been read first, they have a stronger influence on the interpretation of the incoming message than later elements (see Thompson 1985; Hasselgård 1996; Le Draoulec and Péry-Woodley 2003). If we focus particularly on initial elements occurring before the grammatical subject (detached elements), we find elements that may set a discourse frame for the interpretation, as stated by Chafe (1976: 53): "[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."

Orientation is typically the case with the three initial adverbials in example 1. Grammatical subjects may also function as orienters for the rest of the message. In example 1, the first subject establishes the main topic of the entire paragraph (and even of the section).

Initial position is a good starting point for the study of discourse organization: from a linguistic perspective, it allows us to approach the complexity of discourse organization (by taking into account initial adverbials and grammatical subjects) and, from a computational perspective, it makes it possible, thanks to a precise definition of the unit under analysis, to carry out a comparative quantitative analysis based on automatic marking.

3. Configurations of cues for signalling discourse organization

Recent linguistic studies (Jacques and Rebeyrolle 2006; Péry-Woodley 2005) state that discourse organization is signalled by configurations of cues rather than by single markers. Writers and readers have to manage several levels of organization which include not only thematic continuity but also time and space reference, rhetorical articulation and document structure. Discourse cues may simultaneously contribute to several of these interdependent levels. Strong

markers (i.e. lexical expressions absolutely correlated with a designed discourse function) are not as common as are interactions of soft cues. A global view of discourse organization is needed in order to discover these interactions. Such a global view may be achieved by combining three kinds of cues: lexico-syntactic elements, text position and text-type.

3.1. Lexico-syntactic elements

The set of lexico-syntactic elements taken into account corresponds to all the elements occurring in initial position. In this study, initial position is delimited to the preverbal zone of the first sentence of the different textual units (sections, paragraphs, sentences). Figure 2 represents the variety of elements found in this position in French:

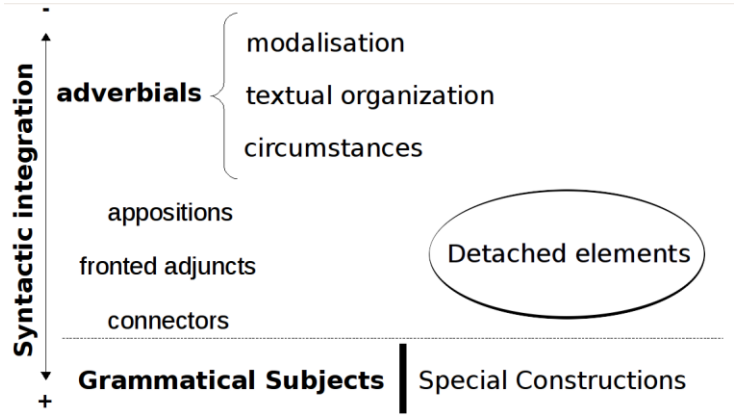

Figure 2. Lexico-Syntactic elements in preverbal zone

A first distinction is made between special and canonical constructions. Special constructions 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 (e.g. presentational constructions) or to indirectly express an evaluation of the process being expressed (e.g. *it*-extraposition). In canonical constructions, the grammatical subject has referential meaning and may constitute the topic of the sentence.

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, separated or not by a comma (e.g. adverbials, appositions, fronted adjuncts). Integrated elements correspond to grammatical subjects.

3.2. Text-type

Textual variation is a feature that must be taken into account in the description of sequentiality. Discourse organization is different depending on whether the text is narrative, descriptive, argumentative, etc. The notion of text-type is defined here following Biber's view: types are determined by linguistic characteristics, in contrast to genres, which are identified on the basis of extra-linguistic parameters such as social function (Biber 1988). In other words, text producers (and readers too) use strategies designed for the different text types. These strategies depend on discourse organization itself rather than on genre. For example, a sub-part of the corpus is composed of texts that have a mono-referential quality in common (they revolve around a single topic). Conversely, there are pluri-referential texts which abound with space and time references. This criterion of text-type characterisation is further explained below, in section 4, which also describes the corpus.

3.3. Text position

The last feature which forms the basis of our methodology is text position. Text position is linked to the level of organization that is marked by the document structure. Orientation and connection processes are likely to vary according to the level of organization: elements in initial position may function differently depending on whether they start 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 the construction of the meaning of a text (for further details, see Power et al. 2003). 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.

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

S1 = section initial position;
P1 = paragraph initial position;
P2 = paragraph internal position.

Through 'playing' with the three text positions in three text-types, relevant configurations of cues will be discovered.

3.4. 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 organization. Although our approach is essentially data-driven, it is strongly shaped by a number of hypotheses which are also tested through this study. The main assumptions are presented in Table 1.

Correlation with	discontinuity	continuity
at experiential level		
figure	Grammatical subjects with a low degree of accessibility, paragraph breaks (P1), headings and new sections (S1)	Grammatical subjects with a high degree of accessibility, paragraph-internal position (P2)
ground	Circumstance adverbials, paragraph breaks (P1), headings and new sections (S1)	Apposition, fronted adjuncts, initial connectors
at rhetorical level		
	Textual adverbials, initial connectors, paragraph breaks (P1), headings and new sections (S1)	Paragraph-internal position (P2)

Table 1. Potential correlations between cues and their contribution in indicating discourse organization

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). Conversely, the introduction of a new referent must be accomplished via elements correlated with low accessibility (e.g. indefinite NP). Although Ariel's accessibility scale is not specifically designed to classify referential expressions in a way that corresponds to the notions of continuity or discontinuity, it provides a model that can be adapted and applied to automatic marking. The adopted correspondences are given in Table 2.

indefinite description or special construction	0	low acc.
proper name without lexical reiteration	1	
long definite description without lexical reiteration	2	
definite description with lexical reiteration or short	3	
reiteration of a proper name ("redenomination")	4	
long demonstrative description without lexical reiteration	5	
demonstrative description with lexical reiteration or short	6	
pronoun or possessive NP	7	high acc.

Table 1: Scale of accessibility adapted from Ariel's Theory of Accessibility (1990)

If we look back on example 1, we can see that in the first sentence, the complete definite description introducing the main topic correlates with a middle-low degree of accessibility (DA = 2). In subsequent sentences, several cues of topical continuity occur: an anaphoric pronoun (DA = 7), a reduced co-referential definite description (DA = 3), a reduced co-referential demonstrative description (DA = 6) and finally a complete demonstrative description (DA = 5).

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 type of text.

(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 the Trattato 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-Ambroise, 1513-1519 *[heading]*
In Rome, where *he* has his lodgings in the Belvedere, *Leonardo* finds himself […]

Examples 1 and 2 are both organized, as regards figure, around a main continuation span and, as regards ground, around different temporal frames. Curiously, the elements in section or paragraph initial position always express 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. These configurations of cues are meant to indicate the organization of the section rather than a coincidence. It is this kind of configuration of cues that the methodology we propose attempts to detect.

4. A data-driven approach

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, all the elements in the preverbal zone are analyzed (and not just a selection of elements for which there are assumptions). After describing the corpus, and setting out the automatic marking, various relevant quantitative analyses will be explained.

4.1. Corpus description

The methodology is applied on a French corpus designed for the study and determined by three choices. The first choice concerns a general category of texts that seem the most relevant ones: long written texts which need a more complex discourse organization than short texts or oral texts. Oral texts strike us 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 the corpus are always over 10 pages in length and divided into sections.

The second choice follows from the first and concerns genre: here expository texts. Expository texts are topic centered, unlike narrative texts where organization is participant and event centered. In expository texts, there is no relation of succession (as happens by default in narratives) or action structure that motivates implicit organization. Moreover, the use of headings and subsections is rare in narratives.

The third choice concerns the parameter of textual variation. The corpus is composed of three sub-corpora representing three text-types distinguished by subject-matter and presentational organization.

a) ATLAS (~205,000 words), composed of 3 descriptive social geography texts;
b) GEOPO (~250,000 words), a collection of 32 argumentative texts in

the domain of international relations;

c) 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 organization is not the norm and never extends to the whole text. GEOPO is more difficult to characterize, with an occasional temporal organization and rather weak topical continuities.

If we count the frequency of nouns in each text, GEOPO and ATLAS show 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, ATLAS has many more basic space adverbials[2] (e.g. *In Europe*) which, moreover, 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.

4.2. Automatic cue tagging

In order to perform an exhaustive analysis without selecting specific cues, all elements appearing in the preverbal zone of every sentence in the corpus are systematically marked. This marking is performed automatically for all the elements carried out in figure 2.

For each of the 23.000 sentences in the corpus, the following features are annotated:

- text position
- sub-corpus
- presence of a detached connective (e.g. *But, And, Nevertheless...*)
- presence of one or more detached elements
- canonical/special syntactic construction

Each detached element (there are 7022 in total) is characterized in terms of:

- part of speech

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

- function (circumstance adverbial, textual adverbial, apposition, etc.)
- semantic meaning for circumstance adverbials (temporal, spatial, notional)

Finally, grammatical subjects are characterized in terms of four properties:

- part of speech
- length (a distinction is made between short NP, consisting of less than four words, and long NP, consisting of more than three words)
- reiteration, i.e. the fact that the NP's head repeats a noun already mentioned in the current section
- degree of accessibility in accordance with the accessibility scale of Ariel (1990) as indicated in Table 2.

These features are automatically detected using a set of regular expressions based on the results of Treetagger and the parser Syntex (Bourigault 2007).

4.3. Measurement of variations

The data are systematically explored in 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 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. For each lexico-syntactic element, the following variations are measured:

a) distributions in each corpus are compared with overall distributions;
b) distributions in each text position are compared with overall distributions.

The significance of variations is given in terms of z-score. We regard as significant positive deviations above +2.5 and negative deviations below -2.5.

The second step consists in measuring:

a) variations in subject position according to the presence of a particular detached element;
b) variations in detached position according to the presence of a particular type of grammatical subject.

These variations are measured in the host sentence and in the following sentence, considering each text-type and each text position.

5. Results and interpretation

Through this exploratory method, a number of results are obtained and presented in Hồ-Đắc (2007). After presenting an overview of the results obtained and their interpretation, the second sub-section illustrates in more detail the method with a step by step account of the study of variations concerning time and space adverbials.

5.1. Organization and text-types

The first set of results presented in figure 3 indicates the general associations showing a significant deviation according to text position. Figure 3 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.

<table>
<tr><th colspan="2">Detached element</th><th colspan="2">Grammatical Subject</th><th></th></tr>
<tr><td>apposition</td><td>Temporal adv.</td><td>Proper name</td><td>Long definite NP</td><td>S1</td></tr>
<tr><td></td><td>Spatial adv.</td><td colspan="2">Lexical reiteration</td><td>P1</td></tr>
<tr><td colspan="2">No detached element</td><td>Pronoun and possessive</td><td>Short definite NP (without reiteration)</td><td>P2</td></tr>
</table>

Figure 3. Significant general associations between lexico-syntactic elements and text-position

If we focus on grammatical subjects, we find well-known associations. Categories that 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:

a) in S1, full definite descriptions and new proper names that may mark discontinuity by introducing a low accessible referent;
b) in P1, lexical reiteration that may be used to emphasize 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 organization.

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.

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, which signal topical continuity, occur significantly more in P1. Only the strongest deviation, concerning space adverbials in ATLAS, is reported with general variations indicated in figure 3.

Table 3 summarizes the different significant variations measured for detached elements according to text-position in each sub-corpus. The same measures for grammatical subjects are indicated in Table 4.

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

Table 3. Detached elements: significant variations according to text position in each text-type.

	S1 ----------> 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

Table 4. Grammatical subjects: significant variations according to text position in each text-type

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 and means that we must pay attention to the correlation between degree of accessibility and signals of (dis)continuity.

Space and the methodological orientation of the present paper prevent us from discussing this point further here or from delving deeper into the detailed analysis of each feature taken into account in this study. We choose to illustrate the methodology we have just outlined by describing the case of space and time adverbials, as they give a good overview of the processes involved in this data-driven approach.

5. An illustration: variations associated with time and space adverbials

5.1. 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 (1466 occ.). Space adverbials are less frequent (7% of all initial elements, 500 occ.). Time adverbials are regularly distributed across text-types: 31% are found in ATLAS, 36% in GEOPO and 34% in PEOPL. This is not the case of space adverbials: 66% are found in ATLAS, 21% in GEOPO and 13% in PEOPL. Figure 4 compares the distribution of these adverbials in each text-type and their overall distribution. The statistical measure employed is z-score.

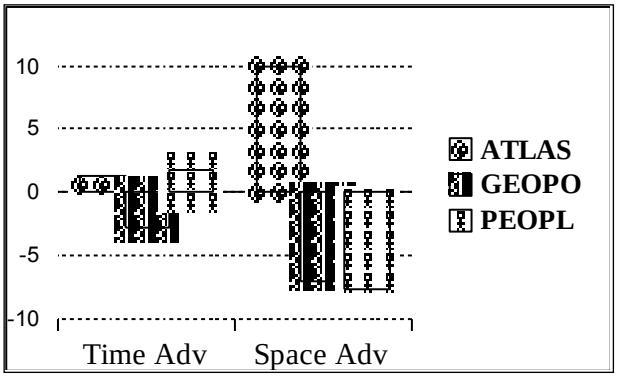

Figure 4. Time and space adverbials: deviations according to text-type

It appears clearly that time adverbials are not specific to one corpus (there is no positive significant deviation in one or more sub-corpora). 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 characterize ATLAS as shown by the strong positive deviation for ATLAS and the two negative deviations for GEOPO and PEOPL.

Variations concerning text positions (S1: section-initial; P1: paragraph-initial; P2: paragraph-internal) are given in Figure 5. Here, the z-score test compares the distribution of elements in each text position with their overall distribution.

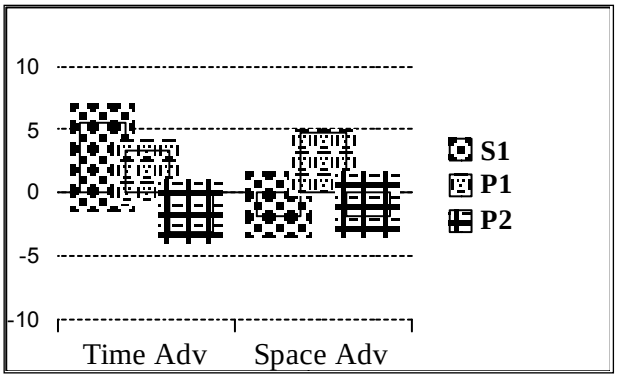

Figure 3. Time and space adverbials: deviations according to text position

Time adverbials can be seen to occur significantly more in S1 and in P1, while space adverbials occur significantly more in P1 only. Conversely, there are significantly fewer time adverbials in intraparagraphic sentences (P2).

The difference between space and time adverbials may be explained in terms of the comparison between local discourse function and global discourse function. Space adverbials are associated with paragraph-initial position but not with section-initial position. Moreover, the deviation is not significant in the case of space adverbials occurring in P2. In contrast to time adverbials, space adverbials are not unlikely to occur in paragraph-internal position. These results may indicate that space adverbials fulfill a more local discourse function than time adverbials. These observations are confirmed by the results below.

Figure 6 displays the results of the same z-score test applied in each sub-corpus. The aim is to discover whether such associations with text positions remain stable across the three text-types.

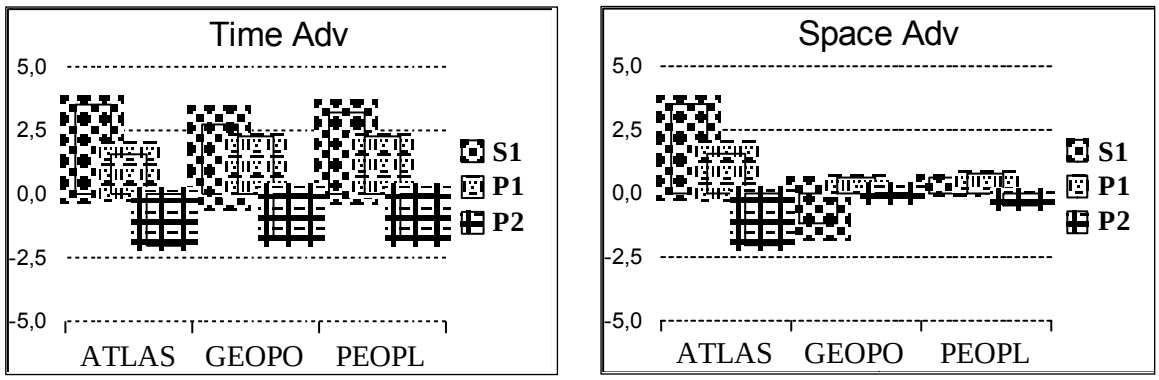

Figure 4. Time and space adverbials: deviations according to text position in each sub-corpus

Deviations associated with time adverbials are found in all three corpora. There are more time adverbials in S1 and P1 and fewer in P2. The z-score in P1 for GEOPO and PEOPL, though not in ATLAS, may be accepted as statistically significant. It is only in comparison with time adverbials that we can find significant deviations affecting the use space adverbials in ATLAS. 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 sections in ATLAS are very short and hierarchically embedded, compared to 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. However, this interpretation must be qualified. Firstly, it is important to take into account text-type (time adverbials appear to be less specific of a text-type than space adverbials). Secondly, this association does not mean that adverbials signal discontinuity on their own, as for example when they appear in a location other than section-initial or paragraph-initial

position. The analysis of the lexico-syntactic environment of adverbials will clarify this last point.

5.2. Step 2: variations in grammatical subject preceded by an adverbial

The second stage of the analysis measures the variations in subject position relative to the presence of an adverbial in detached position. Here, the reader will find only a summary of the most important results (for a complete description, see Hồ-Đắc and Péry-Woodley 2008).

Firstly, the discourse function of space/time adverbials seems to be highly sensitive to text position. Space/time adverbials seem to be good segmentation markers when they occur in S1 or P1. In P2, their discourse function would appear to depend on the textual strategy used in the text: 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 emphasize a topical continuity when there is a shift in the setting (i.e. ground) or in 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. However, the data show 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 (for more details, see Hồ-Đắc and Péry-Woodley 2008).

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 organization 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 not in the least 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 and Péry-Woodley (2003) whereby, 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: *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 organization for the rest of the document structure segment.

6. Conclusion

The data-driven approach presented here provides really interesting new insights. It has proved to be an effective tool for processing 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 organization. It offers new perspectives for the study of discourse organization and enables us to identify the textual characteristics of global organization: for example, the fact that ATLAS is characterized by a strong spatio-temporal organization while the discourse organization in PEOPL is clearly topic centred.

Now concerning advances in the study of discourse organization, the hypothesis concerning the marking of discourse organization has been partially validated. By testing the discourse function of specific lexico-syntactic elements such as time adverbials, this methodology shows that we cannot speak about the structuring power of a lexical marker by itself. It is rather a matter of complex configurations of cues where lexico-syntactic elements play a role. This validation shows also that discourse organization is strongly sensitive to text-type. In this study, the treatment of text-types takes into account the shape of a document and the textual strategies used in it. A promising future direction would be to test the use of the configurations of cues discovered in this study in automatic text-type profiling.

Some aspects of this methodology need further refinement. First, the use of degrees of accessibility to represent the instructional meaning of grammatical subjects must certainly be reassessed. Second, a necessary step to evaluate this methodology will be to apply it to other corpora and other languages in other contexts. We plan to do this in the framework of a project (ARC project – Catholic University of Leuven) aiming to study "The transformation of the relationship with information in multimedia communication" by exploring variations across newspapers on line and on paper.

7. References

Allén, J. (ed.) (1989): *Possible Worlds in Humanities, Arts and Sciences*, Berlin, New-York, Walter de Gruyter.

Ariel, M. (1990): *Accessing noun phrase antecedents*, London, Routledge.

Berry, M. (1995): "Thematic options and success in writing," in M. Ghadessy (ed.), 55-84.

Biber, D., U. Connor and T. A. Upton (2007): *Discourse on the move: using corpus analysis to describe discourse structure*, Amsterdam, John Benjamins.

Biber, D. (1998): *Variation across speech and writing*, Cambridge, Cambridge University Press.

Chafe, W.L. (1976): “Giveness, contrastiveness, definiteness, subjects, topics, and point of view,” in C.N.Li (ed.), 25-55.
Condamines, A. (ed.) (2005): *Sémantique et corpus*, Paris, Hermès.
De Mulder, W. (1997) : « Les démonstratifs: des indices de changement de contexte, ” in N. Flaux, D. Van de Velde and W. de Mulder (eds.), 137-200.
Enkvist, N.E. (1989): “Connexity, interpretability, universes of discourse, and text worlds,” in J. Allén (ed.), 162-186.
Flaux, N., D. van de Velde and W. de Mulder (eds.) (1997) : *Entre général et particulier: les déterminants*, Besançon, Artois Press Université.
Fries, P.H. (1995) : “Themes, methods of development, and texts. In R. Hasan and P. H. Fries (eds.), 317-360.
Ghadessy, M. (ed.) (1995): *Thematic Development in English Texts*, London, Pinter.
Givón, T. (1995): *Functionalism and Grammar*, Amsterdam/Philadelphia: John Benjamins.
Goutsos, D. (1996): “A model of sequential relations in expository text,” *Text*, 16, 4, 501-533.
Gundel, J.K., N. Hedberg and R. Zacharski (1993): “Cognitive status and the form of referring expressions in discourse,” *Language*, 69, 2, 274-307.
Halliday, M.A.K. (1985): *An introduction to Functional Grammar*, London, Edward Arnold.
Halliday, M.A.K. (1970): “Language structure and language function,” in R. Hasan and P. H. Fries (eds.), 140-164.
Hasan, R. and P. H. Fries (eds.) (1970): *New horizons in Linguistics*, Harmondsworth, Penguin.
Hasan, R. and P. H. Fries (eds.) (1995): *On Subject and Theme. A Discourse Functional Perspective*, Amsterdam, John Benjamins.
Hasselgård H. (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.
Hồ-Đắc, L.-M (2007): *Exploration en corpus de la position initiale dans l'organisation du discours*, Unpublished PhD Dissertation, University of Toulouse-UTM, France.
Hồ-Đắc, L.-M. and M.-P. Péry-Woodley (2009): “A data-driven study of temporal adverbials as discourse segmentation markers,” *Discours*, Special issue on Linearisation and Segmentation in Discourse [online].
Jacques M.-P. and J. Rebeyrolle (2006): “Titres et structuration des documents,” *Proceedings of the International Symposium: DISCOURSE and DOCUMENT (ISDD'06)*, 1-12.

Lagerwerf, L., W. Spooren and L. Degand (eds.) (2003): *Determination of Information and Tenor in Texts : MAD 2003*, Amsterdam / Münster, Stichting Neerlandistiek & Nodus Publikationen.

Le Draoulec, A. and M.-P. Péry-Woodley (2003): "Time travel in text: temporal framing in narratives and non-narratives," in L. Lagerwerf, W. Spooren and L. Degand (eds.), 267-275.

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

Li C.N. (ed.) (1976): *Subject and Topic*, New York, Academic Press.

Matthiessen, C. (1995): "THEME as an enabling resource in ideational 'knowledge' construction," in M. Ghadessy (ed.), 20-54.

Péry-Woodley, M.-P. (2005): "Discours, corpus, traitements automatiques," in A. Condamines (ed.), 177-210.

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

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

Sinclair, J. (2004): *Trust the Text: Language Corpus and Discourse*, London, Routledge.

Thompson, S. (1985): "Grammar and written discourse: initial vs. final purpose clauses in English," *Text*, 5, 55-84.

Walker, M.A., A. Joshi and E. Prince (1998): "Centering in naturally occurring discourse: an overview," in M. A. Walker et al. (eds.), 1-28.

Walker, M. A., A. Joshi and E. Prince (eds) (1998): *Centering Theory of Discourse*, Oxford, Calendron Press.