Linguistic Markers of Lexical and Textual Relations in Technical Documents

Anne Condamines, Marie-Paule Péry-Woodley

Linguistic markers of lexical and textual relations in technical documents

Anne Condamines & Marie-Paule Péry-Woodley

(anne.condamines, pery)@univ-tlse2.fr

Equipe de Recherche en Syntaxe et Sémantique

CNRS et Université Toulouse Le Mirail

5 allées Antonio Machado

F-31058 Toulouse cedex

This chapter proposes a number of linguistic "handles" for the description of technical documents, at a lexical level (terminology) and at a textual level (discourse coherence). Examples are given of uses of such insights in document production and management, in particular *via* document engineering systems.

We provide a number of linguistic "handles" for the description of technical documents. Such insights into the "inner workings" of texts may be harnessed in various ways in the production and management of technical documents; we show some applications in document engineering, in systems designed to facilitate access to information. Our focus is on surface markers, i.e. observable text features identified through corpus analysis, signalling the kind of relations between lexical items used in building terminologies (such as *generic/specific*, see section 1), or relations between text segments involved in discourse coherence (such as *theme*, or *rhetorical relations*, see section 2). We insist on the relevance of the notion of *genre* when working with technical documents, and on the genre-dependent nature of our linguistic markers.

Our objective in this chapter is to provide a number of precise, theoretically motivated and descriptively relevant "handles" for the linguistic analysis of lexical and organisational aspects of technical documents. By "handles" we mean observable text features which can be associated with specific functions in the document – whether marking semantic relations between lexical items (section 1) or signalling discourse organisation (section 2).

Though concerned with two distinct types of semantics, the two sections share a number of methodological options: they focus on studies concerned with the precise analysis of linguistic realisations (surface forms) in attested texts (as opposed to made-up or experimental texts); a starting hypothesis for both is that the markers they are attempting to circumscribe are likely to be sensitive to genre, and that generalisations cannot safely extend beyond the genres that have actually been described.

1. Using linguistic markers to build terminologies for the management of technical documents.

1.1. The problem

This first section explains how terminologies can be used for documentation engineering.

1.1.1. Documentation management

Documentation management has become an important issue for companies where each manufactured product is accompanied by numerous documents necessary to build, maintain and market this product.

For example, it is often stated that the documentation for an aircraft would fill this aircraft or also that the documentation for a space project amounts to 150 000 pages of paper.

This situation is somewhat contradictory. On the one hand, we have technical objects that are highly sophisticated and considered to be extremely reliable, and on the other hand we have documents written in natural language with all the inherent difficulties that implies: ambiguity, polysemy, etc. One possibility to limit these difficulties is to try to standardise document authoring and this is sometimes mandatory in fields such as aeronautics (AECMA norms), the goal being to establish rules in order to guide technical writers in aeronautics (http://www.aecma.org/Publications.htm).

The most common norms concern terminologies. Most of the time, terminologies are built by experts in a given field who decide to establish definitions within this field. For this purpose,

terminologists meet writers'needs when they build thesauri and some firms have tried to define their own thesaurus (see for example the Nasa's thesaurus which contains 13 000 words, 9 000 acronyms and 10 000 definitions).

In spite of the interest of these standards, it is clear that they are not much used. Sometimes, writers do not even know that they exist. One of the main problems is that these standards are established by official bodies with the aim of covering an entire field. However, the proposed terms do not always correspond to the ones actually used by a particular company.

Some years ago, the problem of documentation management took on a new perspective when it was examined by knowledge engineers. This has led to major changes with regard to the problems of terminology.

1.1.2 Terminologies and knowledge engineering

During the last 15 years, there has been a significant evolution in artificial intelligence. The important point concerned the need to consider that it is not possible to build tools to take the place of humans reasoning but only tools to help humans in their reasoning. Then the development of knowledge engineering began. The most important element in this new perspective was the distinction between knowledge systems on the one hand and reasoning systems on the other (Clancey, 1993).

Initially built on the basis of interviews with experts in a given field, these knowledge-based systems are now very often built on the basis of document analysis; this new perspective is particularly relevant for documentation management as these systems can also be used for this purpose.

The knowledge module uses a kind of representation called an *ontology* which presents clear similarities with terminologies as it is formed by nodes linked by relations, both labelled by lexical elements. But ontologies are different in an important respect: they must be formalised in order to be integrated into reasoning systems (Gruber, 1991). An ontology may be defined as:

"a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents" (Gruber, 1993).

With the possibilities of Natural Language Processing (NLP), it then became possible to envisage the design of tools adapted to the reality of language use within a specific firm's

documentation in order to build terminologies or ontologies. In return, these terminologies can be integrated into tools devoted to document management.

Many NLP tools have been designed with the purpose of helping to create terminologies or ontologies from texts. All these new perspectives have boosted classical documentation and terminology and given birth to new kinds of studies.

1.1.3 Development of a textual terminology

The first terminologist's theory (Eugen Wüster) was based on the idea that it is fundamental to standardise terminologies in order to allow good communication between speakers in a given field (Sager, 1990), (Cabré, 1999) (Rey, 1995). It is easy to understand the idea underlying such a point of view: standardisation can be seen as the only way to guarantee transparent communication between humans in a field. But this point of view does not consider the reality of language which is always evolving.

In the early 1990s, several interdisciplinary teams identified the relationships between the aims of terminology and knowledge engineering, and a joint reflection process was initiated (Skuce & Meyer, 1991).

One of the main consequences of this new goal was that terminology and documentation in companies were considered to be important activities and, in return, they were forced to re-examine their own methods and their situation among other disciplines.

In terminology, a new field appeared that examined how it is possible to build terminologies from texts: textual terminology (Pearson, 1998).

With such an evolution, terminology has come into the spotlight as a linguistic issue, or more precisely, a corpus linguistics issue. The main point in corpus linguistic studies is to try to take into account variations in use and identify regularities among these variations in order to build a system that can explain them. Textual terminology has a very similar goal: taking into account real documents in order to create a system of terms, i.e. a representation of the content of these documents in a relational form (terms linked by semantic relations).

This process must take into account how the terminology will be used. There may be variations in the building of a relational representation according to the aim of the project. Therefore, building a terminology from a corpus supposes an interpretation.

The linguist'role consists in explaining how it is possible to construct a terminology, more precisely to describe on which linguistic elements this construction can be performed.

Building a terminology from texts (or an ontology, as during the first step, the perspectives are very similar) requires the identification of terms and of the relations between them.

Different ways of identifying terms have been developed, both from linguistic and NLP viewpoints (Cabré, Estopa & Vivaldi, 2000) but, in this paper, we prefer to focus on the problem of the identification of relations between terms.

From a linguistic point of view, it is really interesting to examine how it is possible to identify relations between terms, especially by using relation patterns. These relation patterns constitute a kind of handles. (Meyer, 2001, p.290) considers that there are three kinds of conceptual relation patterns (named knoweldge patterns by Meyer) : lexical patterns ("involving one or more specific lexical items"), grammatical patterns, paralinguistic patterns ("they include punctuation, as well as various elements of the general structure of a text"). The author precises that these patterns are "complex in their nature, and in the way they can be relized in text" : they are sometimes unpredictable, polysemic, and/or domain dependant. In the next part, we are going to describe the role of these patterns in the identification of conceptual relations in texts.

1.2 The role of patterns in the identification of relations between terms

Before the development of textual terminology, linguists had already identified lexico-syntactic elements "expressing" semantic relations such as [all N1 except N2] for hyperonymy (generic/specific relation) (e.g: from *all flowers except roses* we can say that roses (specific) are a kind of flower (generic). This was the case for Cruse who named these elements "diagnostic frames" (Cruse, 1986) or Lyons who named them "formulae" (Lyons, 1977). The study of these patterns has been developed over the past 15 years, specifically to design tools (Hearst, 1992). It is also very interesting to explore this field from a linguistic point of view both because it can constitute a way to understand how variations may be taken into account in corpus linguistic methodologies and because it may help to improve NLP tools.

In this case, the problem is to explain why relation patterns do not appear in all texts and/or in the same way. In other words, it is necessary to understand how the nature of a corpus can play a role within the conceptual relation/pattern pair. This point is really important in order to improve information spotting specifically in a NLP perspective. Concerning technical documentation, it is important to evaluate if a determined pattern may be present or not considering the nature of this text (or set of texts).

Three kinds of dependences between corpus and patterns have been identified (Condamines, 2002).

1.2.1 Weak dependence

Some patterns seem able to appear in any text. This is the case for some hyperonymic (generic/specific relation, see above) or meronymic (part-of relation) patterns such as:

[N1 comprises N2, (N3) and N4] as in:

> *[1] The house comprises a living-room and two bedrooms* (*living-room* and *bedroom* are parts of *house*).

These patterns are the ones generally proposed spontaneously and used in NLP tools. Even if the relation/pattern seems very strong, some difficulties should be noted with these patterns. First of all, these patterns may be polysemic. For example *comme* (as) in French may be associated either with a hyperonymic relation or a comparison relation.

Another difficulty (and this is the case with all patterns) is that it may be difficult to determine if the speaker using conceptual relation patterns is expressing his/her own point of view or if he/she is presenting the point of view of a group of speakers. In the perspective of terminology construction, only the second case needs to be modelled because such models must be acceptable and reusable within collective tasks.

Finally, despite their strong link with a relation these patterns do not always appear because the relations are not always used within texts. This may be surprising when general relations such as hyperonymy (generic/specific) and meronymy (part of) are considered as very significant for structuring a field. Nevertheless this case was found in a corpus from Matra Marconi Space built from specifications in the field of satellite simulations. There were no classical patterns of hyperonymy (such as the one presented above) because hyperonymy was not expressed. In this very specialised field, experts do not need to explain concepts; they use them under the hypothesis that they will be understood by the readers. So concepts are never defined, i.e. never situated in relation with other concepts.

What can be retained from such an experience is that the corpus must be built according to the goal of the study. To build terminologies and in particular identify semantic relations, it is important to explore texts written by experts for less expert readers, in other words, texts with some kind of didactic intention.

1.2.2 Complete dependence

In some corpora, some unpredictable structures play the role of patterns for some relations. It is impossible to propose them spontaneously (Meyer, 2001). We found such a structure in a corpus from EDF (Electricité de France) concerning specifications and the writing of documents in computer science (Condamines A. & Rebeyrolle J., 2001). In this corpus,

written by several experts but with the same purpose, the pattern for the relation of condition was:

[(phase, stage) or nominalisation + (when, as soon as) + passive V]

[2]La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.

The component integration phase can begin when all the software elements have been coded.

This example should be understood as:

Software elements must have been coded for the component integration phase to begin.

Therefore, the relation expressed is a condition relation and the pattern was very productive as for each of its 11 occurrences, the condition relation was present.

These patterns are very difficult to identify and only a fine-grained analysis brings them to light.

1.2.3 Dependence in terms of text genre

Sometimes, regularities of expression are not specific to a particular corpus but rather to a corpus to the extent that it is "representative" of other corpora with identical extra-linguistic and linguistic characteristics, i.e. when it is possible to identify the text genre to which the corpus belongs. This notion of genre is very ancient (Aristotle proposed to distinguish three kinds of genre: lyrical, epic and dramatic). In the middle of the 20th century, this notion was reactivated by two movements: one was a Russian movement which anchored this approach in a dialogical point of view (Todorov, 1984) and the second was an American movement, within sociolinguistics, which considered that text genre was the only way to take into account both situational and language regularities (Firth, 1969).

This notion of genre is used within corpus linguistics, which has a fundamental need to organise corpora in order to try to explain variations (Bahtia, 1993), (Swales, 1990). But this notion of text genre is not easy to determine because its relevance may vary according to the element under study.

Concerning relation patterns, text genre is very often relevant but it does not affect all patterns in the same way.

Let us examine two examples of patterns. The first one concerns the preposition *chez.* In some cases, this preposition occurs in sentences where a meronymic relation can be identified:

[3] Chez les colobinés, le nez fait saillie sur la lèvre supérieure.

With the colobines, the nose juts out over the upper lip (there is a meronymic (part-of)

relation between nose and colobines).

This phenomenon appears in particular in some specific texts: those that belong to natural science (only biology and zoology excluding geology) and are didactic (e.g. encyclopaedia about natural sciences). This means that, in texts belonging to such text genres, the probability for sentences with *chez* to be interpreted as a meronymic relation is high.

Beyond this quantitative aspect, it is interesting to study different examples containing this preposition to understand how the meronymic interpretation is possible. The point is that this preposition introduces a referent (either at the beginning, the middle or the end of the sentence) and in didactic natural science texts, what is said about this referent (animal or plant) concerns very often their anatomy or composition. Quantitative results show that anatomical information is more present than other kinds of information (habitat, feeding, reproduction etc.), around 50% of the occurrences. So it is not really the case that *chez* is a pattern for the meronymic relation: it does not intrinsically contain this sense. But, from a computational point of view, *chez* may be used to identify structures where a meronymic relation occurs in didactic texts of natural science. This case is very interesting because it shows that the linguistic and computational points of view are not always equivalent.

Another pattern has been studied with the same hypothesis concerning genre. It is the nominal anaphora. It is well known that in some cases, there is a generic (or hyperonymic) relation between a noun in anaphoric position and its antecedent (Cornish, 1986):

[4] Contrôle de la complétude des modification effectuées. Cette activité est du ressort du Responsable Développement. (activité est un générique pour contrôle).

Completeness check for the modifications made. This activity is the responsibility of the Development Manager.(activity is a generic of check)

However, the relation may also be of another nature; for example, the anaphoric noun may be a synonym or a nominalised form of a verb (*This guide proposes […] this proposition*).

The intuitive hypothesis was that this hyperonymic relation could be very frequent in specialised handbooks since it is said that relations within specialised fields are extremely stable.

We have studied nominal anaphora in three specialised handbooks and compared these results with texts belonging to other text genres (literary and journalistic). The results did not confirm the hypothesis. The frequency of hyperonymic relations in the handbooks compared with other genres was not very significant except for one of the technical handbooks. In the other four texts, the hyperonymic relation appears only between 15 and 30 % of all occurrences. It

is therefore not possible to consider that nominal anaphora can constitute a pattern for the hyperonymic relation in handbooks; this did not confirm the initial hypothesis. Nevertheless, fine-grained analysis shows that there are differences between technical handbooks and other texts. Several characteristics allow us to say that anaphoric nouns in handbooks are more often classifiers than in texts of another genre. This means that these nouns may be considered as the top-nodes (generics) of possible classes. Thus, even if the nominal anaphora cannot be considered as a hyperonymic pattern, it can be considered as a hyperonym marker: it is not the relation itself which is identified, but only one element of this relation, hyperonym (specific). So, it would therefore be necessary to use other patterns in order to identify specific terms corresponding to generic terms identified by the anaphora pattern.

These two examples show that genre may be relevant for the description of relation patterns but it is very difficult to determine in what cases it will be relevant. Only large studies of real corpora would make it possible to predict usage.

More and more studies are trying to understand how genre can be used to determine and anticipate variations within texts, especially in texts from restricted fields (Trosborg, 2000).
For documentation engineering, it is clear that this notion could also be relevant. Even probably reduced in such specialised contexts, the different genres and the dividing lines between them are not easy to determine and this will probably constitute an important issue in the next years.

2. Three levels of text organisation in technical documents: models and markers

A defining characteristic of texts is that they form a whole, they show connectedness. Hence the importance of the notion of *coherence* in text and discourse linguistics. Coherence is most usefully seen as a mental phenomenon, rather than an inherent property of the text. To quote Sanders and Spooren: "*Language users establish coherence by relating the different information units in the text*" (2001, p. 7). "Relating information" is what discourse is about, whether one looks at production or comprehension[1]. A number of hypotheses have been put forward – elaborated to a lesser or greater extent into models – which can be seen as different – often complementary – takes on the complex notion of coherence. For the textual analyst

[1] See Madrid and Cañas, this volume, on cognitive taks involved in text comprehension and in the construction of a coherent interpretation in the specific case of hypertext

concerned with practical issues of communicative efficiency, they provide the necessary theoretical foundations to interpret and integrate observational data. A strong hypothesis here is that variations in wording, the presence or absence of certain "markers" – lexico-grammatical or visual – have an impact on meaning, i.e. they are used as signals in the construction of the interpretation model.

As a presentational device, this section on the textual organisation of documents is structured in terms of levels of granularity, from the proposition to the whole document. At the finest level, we will look at how contextual and co-textual factors can influence wording and word order in the proposition or sentence (*information structure*). At the next level, the focus will be on relations of connection between these basic units (*rhetorical structure*), before moving on to envisaging the document more globally, and also as a visual object (*document structure* or *text architecture*).

The descriptive studies presented below mostly concern technical documents, as the specific realisations are sensitive to parameters of genre; the models referred to, however, may have been developed with little reference to genre[i].

2.1. Information structure and information packaging

The notion of *information structure* belongs at the fine granularity level, despite the potentially deceptive, apparently all-encompassing term. The more expressive denomination "information *packaging*" was coined by Chafe (1976) to refer to the effects of a combination of factors – mostly to do with prior knowledge and cognitive state – which have a major role in shaping utterances: "*The kind of phenomena at issue here (...) have to do primarily with how the message is sent and only secondarily with the message itself, just as the packaging of toothpaste can affect sales in partial independence of the quality of the toothpaste inside*" (Chafe, 1976, p.28). Chafe's intuition has been elaborated upon over time, and several parameters have been distinguished, amongst which *aboutness* – whether a particular referent or entity is what the sentence is about; *givenness* – whether it is already present in the discourse; *activation* – whether it is the hearer's current focus of consciousness. These parameters influence linguistic choices in ways which differ from language to language, affecting stress, word order and syntactic choices. For instance, different configurations of these parameters lay behind the contrast between "The pipes are RUSTY" and "The PIPES are rusty"[ii] corresponding respectively to the questions: "What about the pipes? In what condition are they?" (the entity referred to by "the pipes" is given in the question) and "Why

does the water from the tap come out brown?" (the entity "the pipes" is new) Another example is the contrasting placement of a circumstantial adjunct: "Tomorrow, John is leaving" vs. "John is leaving tomorrow", adequate responses respectively to "What's happening tomorrow," and "When is John leaving?". The question-answer minimal pairs are of course a somewhat contrived device for introducing context; what must be stressed is that "Tomorrow, John is leaving" in answer to the second question, though syntactically correct, is seriously flawed from a discourse point of view. Another angle on this is to look at the choice of referring expression (e.g. pronoun vs. definite expression) as a signal given to the reader as to the degree of mental accessibility of a piece of (given) information: for instance the use of the pronoun "he" or "she" implies a highly accessible referent, whereas a definite expression ("June's friend") indicates a relatively low degree of accessibility. For a detailed account of information structure, see Lambrecht (1994), of the notions of theme and topic, Gomez-González (2001), of accessibility theory, Ariel (2001).

Importantly for workplace communication, therefore, writers (and speakers) have a choice of different linguistic realisations for the same propositional content, and must at every step shape their utterances in accordance with current assumptions about the readers'(or hearers') cognitive state at this point in the discourse, as well as with their discourse intentions. In turn, the specific shape of the utterance works as a set of signals, or instructions, to the reader. In a ground-breaking article, Grosz and Sidner (1986) propose a model for relating "attentional state" and speakers' intentions in a task oriented dialogue. They look at how entities come in and out of focus, and how processing decisions at a local level are constrained by the textual form of an expression, as determination (e.g. definite or indefinite) or syntactic function (e.g. subject or object) make it a more or less likely candidate for attentional focus. This study opened the way for an important dynamic model relating discourse intentions and attentional states (changes in focus of attention): Centering Theory (cf. Walker, Joshi and Prince, 1998).
The relations of specific "packagings" to particular configurations of information and cognitive parameters are of obvious interest to linguists and applied linguists (Davison, 1984; Foley & Van Valin, 1985; Fries, 1995; Gundel, Hedberg & Zacharski, 1993; Prince, 1981; Virtanen, 1992a, *inter alia*). In the wake of Clark and Haviland's "given-new contract" (1977) a number of psycholinguistic studies have shown the negative impact on comprehension of text disrespectful of information structure, for instance violating the given (information) before new (information) principle. On the basis of these linguistic and psycholinguistic

studies, the information structure approach is clearly highly relevant to the study of the production and comprehension of professional documents.

A number of linguistic devices are associated with information packaging choices, as they allow a reshuffling of elements away from the canonical word order: passivation, clefting, topicalisation, and differential positioning of adjuncts. Though they affect the ordering of elements within the sentence, they reflect contextual constraints, where pragmatics touches on syntax, and they can play a role in the development of larger textual units. With regard to technical and more specifically procedural texts, a particular question concerns the ordering of action pairs when an instruction is given in relation to a purpose, as in the following examples borrowed from Delin, Hartley and Scott (1996):

[5] *a. In order to turn on the light, flick the switch.*

b. Flick the switch in order to turn on the light.

Thompson (1985) suggests a strong functional contrast between these two positions: whereas final purpose clauses (as in [5]b.) have a purely local role, merely stating the purpose for which the action named in the main clause is undertaken, initial purpose clauses (as in [5]a.) "guide the reader's attention […] by naming a problem which arises from expectations created by the text or inferences from it, to which the following material, often consisting of many sentences, provides a solution" (1985, p. 67). A series of initial purpose clauses, each extending their scope over a number of instructions, can structure a passage, reflecting a particular text-building strategy (cf. Péry-Woodley, 2001; Virtanen, 1992a and b). Delin *et al.* (1996) propose a framework for the contrastive analysis of such choices (in English and French instructional texts) based on the semantic relations of *generation* and *enablement* (cf. Goldman, 1970).[iii] This text-organising role observed in the case of purpose clauses placed in initial position also applies to other detached adjuncts, such as time or place adjuncts (e.g. "for the first thirty minutes"), or praxeologic adjuncts (e.g. "in biochemistry"), a behaviour which has been studied under the term of *discourse framing* (Charolles, 1997; Charolles, Le Draoulec, Péry-Woodley & Sarda, 2005; Péry-Woodley, 2005).

It appears quite clearly already that the level of granularity we started with, the sentence, cannot possibly be seen as self-contained, as many aspects of linguistic realisation at sentence level are constrained by higher levels of textual organisation, and in turn influence interpretation so far, and the expectations upon which further interpretation will proceed. The relations between propositions (purpose clause and main clause) just considered provide a

transition with the next level of granularity as they can apply between sentences as well as within sentences.

2.2. Rhetorical structure: connecting text spans in a meaningful way

Coherence relations, discourse relations, rhetorical relations – different terms from different models for the meaning relations which connect text segments (such as Cause-Consequence, Problem-Solution). In the previous section, we looked at purpose clauses with respect to their position in the sentence, and the role this position confers on them, a role which may extend over a wider text segment. Such subordinate clauses, whichever position they occupy, are one possible way of materialising in text a semantic or rhetorical relation between segments – here a purpose relation between two propositions. In [5], the relation is made explicit *via* a particular syntactic construction, but authors agree that relations may be realised in diverse ways, including implicitly. In [6] below, an extract from a software manual, the appearance of the dialog box (second sentence) is likely to be interpreted as resulting form the action instructed in the first sentence:

[6] From the Project menu, choose Components. The Components dialog box appears.

Yet there is no cue to a relation between these text contents beyond mere juxtaposition, which iconically suggests temporal succession. The level of explicitness of relations is linked to writer's assumptions about the reader (e.g. regarding competence level) and the situation (e.g. greater necessity to guide the reader through explicit use of markers in highly technical or risky situations).

There are a number of models and a wealth of studies of discourse relations, with a general consensus on their vital importance in the comprehension process (see Bateman and Rondhuis (1997) for a review covering several models). Among these, Rhetorical Structure Theory (RST, Mann & Thompson, 1988, 1992) [iv] has over the years become a sort of reference model, widely known and used in different communities (descriptive, computational and psycho-linguistics). RST posits a basic asymmetry between the members of most relations: thus in [5] the purpose ("to turn on the light") is seen as a *satellite* in relation to the action instructed, the *nucleus* ("flick the switch"), an asymmetry conveyed in this example through syntactic status (subordinate vs. main). RST relations apply recursively, with text spans resulting from the application of a relation entering into further relations and so on; at the highest level of representation, if a text lends itself to a coherent reading, it should be possible to represent it by a single overarching relation. This combination of recursive span

construction and asymmetrical informational status of satellite and nucleus can be exploited to select informationally richer text spans (see Marcu (2001) for an implementation in an automatic summarization system).

In short, discourse relations are seen by most authors as serving a twofold text building role: they connect segments *via* semantic and/or rhetorical links, and they create a hierarchy of segments, some appearing as subordinate to others. The importance of these functions for efficient technical writing is clear. A number of specific studies of discourse relations in instructional and explanatory texts have been conducted, mostly with a view to computational applications such as automatic text generation (Grote, 1999; Scott, Delin & Hartley, 1998; Vander Linden & Martin, 1995). A major problem for automatic generation is the variability in the markers used to express relations, which is equally a problem for human text production and comprehension. Much research in descriptive and computational linguistics has focused on identifying cues associated to particular discourse relations (Knott & Sanders, 1998; Redeker, 1991 *inter alia*), while psycholinguistic studies have researched the impact of their presence or absence on comprehension (Degand, Lefèvre & Bestgen, 1999; Sanders & Noordman, 2000; Townsend, 1997), or taken the RST account of relations as a tool for studying the writing process (Torrance & Bouayad-Agha, 2001).

2.3. Document structure: the linguistic nature of layout

The most immediately obvious form of document organisation is its visual – graphical – structure: a long document is typically organised in chapters and sections – headed by titles or headings – then in paragraphs and *text-sentences* (Nunberg, 1990), within which various further *textual objects* (Virbel, 1989) stand out through contrasting disposition (e.g. indentation) or typography (e.g; bold face). Luc and Virbel (2001) stress that a graphical token cannot be devoid of visual properties – shape, size, colour – and must be interpreted spatially in relation to other tokens; these properties cannot be envisaged as a simple coding of an already constituted message, they have to be seen as playing a part in the realisation of the writer's intentions. Though layout issues have generally been overlooked in linguistic approaches to discourse, a number of authors have been keen to study the specific potentialities of written language linked to its visual realisation: Nunberg (1990) analyses punctuation as manifesting a coherent linguistic subsystem (*text-grammar*) coexisting with what he calls the system of *lexical grammar*; a distinction picked up and extended by Power,

Scott and Bouayad-Agha (2003) in their study of *document structure* in patient information leaflets. Virbel calls *text architecture* the text structures which are realised *via* the physical page layout (see Luc & Virbel (2001) for a synthetic presentation of the "Model of Text Architecture"). These authors (along with Delin, Bateman & Allen, 2002), though they may differ over several points, agree on some fundamental principles and points of interest:

a) they stress the need to distinguish between the concrete realisation of the graphical form of text (typography, punctuation, disposition) and an abstract structure, diversely called *document structure* (Power *et al.), text architecture* (Virbel), *layout structure* (Delin *et al.*); this abstract structure interacts with choices in wording, and is therefore an aspect of the linguistic realisation of discourse acts.

b) they focus on the interaction between this abstract document structure and rhetorical structure (propositional meanings and their semantico-pragmatic relations). Both Luc and Virbel (2001) and Power *et al.* (2003) show that the two structures are distinct and not necessarily isomorphic. They both address the problem for RST, whose representations are based on relations between text spans (document structure, architecture) when in fact the relations are between text meanings.

Studies of the impact of layout on written text comprehension are of obvious relevance in the workplace context (on instructional texts, cf. Garcia-Debanc (Ed.), 2001). A number of recent studies approach document structure in its interaction with other discourse structures: Bouayad-Agha, Scott and Power (2001) look at the impact of layout on the interpretation of referring expressions, Luc, Mojahid, Virbel, Garcia-Debanc & Péry-Woodley (1999) focus on a structure of particular interest in this perspective: enumeration. The signalling of enumeration can be placed on a continuum from purely discursive (linear form with lexical markers – e.g. "first…, second…") to purely visual (vertical disposition with indentation and bullet points). Linguistic and psycholinguistic studies of enumerations have turned them into a sort of test case for a view of rhetorical and document structure as separate, interacting, types of structure (Luc *et al.,* 1999; Luc, Mojahid, Péry-Woodley & Virbel, 2000; Power *et al.*, 2003; Garcia-Debanc & Grandaty, 2001; Carrio, 2006). It seems important to take further the understanding of a structure which appears to be a fundamental way of organising text, and is a major device in new document forms (cf. homepages of most websites).

Technical documents in the workplace: linguistic studies and document engineering

We have presented approaches to the analysis of linguistic aspects of documents which strike us as being essential keys for the study of their production and comprehension in the workplace. These approaches typically focus on surface features which, in a particular genre, may signal a semantic or textual function. Two types of descriptive "handles" have been described.

Section 1 looked at surface patterns signalling lexical/conceptual relations, i.e. linguistic features (lexical or grammatical) that can be used in order to identify semantic relations such as generic/specific, part of, is cause of... Our focus has been the link between these patterns and the nature of the corpus: the fact that the probability of occurrence of a pattern, as well as its interpretation, are corpus-dependent. In order to describe and explain these variations, we call upon the notion of text genre, which allows us to take into account both extra-linguistic and linguistic features. Variations in extra-linguistic parameters lead to variations in wording. Text genre thus constitutes a way of anchoring linguistic phenomena in sociological contexts and of taking into account the reality of linguistic usage. We also saw that the degree of genre dependence is highly variable, with some patterns totally genre dependent and others applying across genres.

In Section 2, we also looked at markers and relations, but this time our focus was text construction, rather than relations between lexical items or concepts. Accordingly, the markers considered included visual features of documents (typography, disposition) as well as lexico-grammatical expressions. There is a clear cognitive dimension to the research presented, concerned as it is with the textual basis for the construction of an interpretation by readers. We considered three interrelated aspects of text construction which may be seen as vital for comprehension, and therefore have to be carefully "encoded" by the writer: information packaging - given vs. new information, theme vs. rheme; relations between text segments – what is said in segment B is meant to be understood e.g. as the consequence/result of what is said in segment A; and finally text architecture as the abstract structure underlying the graphical realisation of documents.

These linguistic studies take on particular relevance in the current technological context, with the spread of digitised documents leading to the development of new modes of production, access and management of documents in work situations. They mostly have to do with information overload and the need to access selected information efficiently. We propose to give a brief overview of some applications in language and document engineering where the

identification of linguistic markers is important: information retrieval, information extraction, automatic summarization.

The term "information retrieval" designates a process which aims to identify, in a textual database, documents corresponding to a query. In order to reduce silence, the search may be extended to other semantically linked key words. For query extension, information retrieval systems may use linguistic resources constructed on a semantic basis, i.e, terms and conceptual relations; these resources may be elaborated using patterns such as the ones presented in this chapter.

In the case of information extraction, the goal is to determine which text elements correspond to categories of information that have been identified as relevant for the domain. For example, in the case of dispatches, the system has to identify what happened, where, when, why and so on. The linguistic approach is to characterise linguistic structures corresponding to these categories of information in order to retrieve this information as reliably as possible. Obviously, descriptions are guided by the fact that linguistics regularities appear according to text genre. The markers described in this chapter belong to such regularities: they take into account lexical and grammatical elements but also their place in the discourse.

Automatic summarization, or document synthesis, aims to produce a shortened version of a document while retaining the most important points of the text. Given current technological limitations – no computer can "understand" a document – most systems rely on extraction techniques, i.e. the selection of "key" text segments which are extracted and assembled to form the summary. The selection is based on a composite score, with a major lexical statistics component, to which are then applied various weightings. This weighting stage is where markers of the type described above may be called upon, as for instance so-called "cue phrases" which signal segments with a specific rhetorical function (e.g. "*In summary*", "*In conclusion*", see *inter alia* Mani (2001); Minel (2003) for a recent account). Marcu (2001) proposes a method to identify discourse units on the basis of connectives and punctuation, then produce a complete rhetorical tree according to RST (cf. above), which can then be pruned of some of its satellites to retain the "most important" information. Other approaches use the rhetorical conventions of certain genres (e.g. scientific papers) to help the user find "zones" of text with a particular function in the argument (Teufel & Moens, 1999).

This last mention introduces a new actor in document synthesis: the document user. Initially, most approaches took for granted that there were objectively "more important" text segments. New systems are now designed to take into account the user's aim in consulting a document. Various levels of interactivity are introduced, which blur the boundaries between applications:

information extraction can be seen as a form of summarization where the user determines in advance what information is wanted from the text base. Question-answering systems constitute a totally user-centred form of consultation, which takes no account of writer's purpose. Document browsing systems may be seen as both text- and user-sensitive: they are interactive systems designed to help readers find, in a long document or in a series of documents, segments which are relevant with respect to a query (Minel, 2003; Bilhaut et al, 2003). They use various discourse markers (e.g. frame introducers) together with other techniques and aim to provide sophisticated display functions, so as to overcome the disadvantages of on-screen reading

These applications constitute an important domain for natural language processing. But, from a linguistic point of view all the issues have not been sufficiently envisaged. Thus, it will be necessary to develop linguistic analysis to evaluate the possibilities of improve such systems but also to better understand specificities of technical documents.

References

Ariel, M. (2001). Accessibility theory: an overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29-87). Amsterdam: John Benjamins.

Bahtia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.

Bateman, J. & Rondhuis, K.J. (1997). "Coherence relations": Towards a general specification. *Discourse Processes*, *24*(1), 3-50.

Bouayad-Agha, N., Scott, D., & Power, R. (2001). The influence of layout on the interpretation of referring expressions. In *Proceedings of 4th International Multidisciplinary Approaches to Discourse (MAD).* Workshop. Ittre, Belgium. 133-141

Cabré, M.-T. (1999). *Terminology: Theory, methods and applications*. Amsterdam: John Benjamins.

Cabré, M.-T., Estopa R., & Vivaldi J. (2000). Automatic Term Detection: a Review of current systems. In D. Bourigault, C. Jacquemin, & M.-Cl. L'homme (Eds.), *Recent advances in computational terminology* (pp. 53-87). Amsterdam: John Benjamins.

Chafe, W. L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. In C. N. Li (Ed.), *Subject and Topic,* (25-56), New-York: Academic Press.

Charolles, M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces. *Cahier de Recherche Linguistique* 6, LANDISCO, URA-CNRS 1035 Université Nancy 2. 1-73.

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

Clancey, W. (1993). The knowledge level reinterpreted: Modelling socio-technical systems. In K. Ford & J. Bradshaw (Eds.), *International Journal of Intelligent Systems*. Special Issue on Knowledge Acquisition as Modelling. *8-1*, 33-50.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.) *Discourse Production and comprehension*. Norwood, New Jersey: Ablex. 1-40

Condamines, A. (2002). Corpus Analysis and Conceptual Relation Patterns. *Terminology*, *8-1*, 141-162.

Condamines, A., Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a terminological knowledge base (CTKB): method and results. D. Bourigault, M.C. L'homme, C. Jacquemin (eds), *Recent advances in computational terminology.* (pp.127-148). Amsterdam: John Benjamins.

Cornish, F. (1986). *Anaphoric relations in English and French, a discourse perspective*. London: Croom Helm.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, *60*(4), 797-846.

Degand, L., Lefèvre, N., & Bestgen, Y. (1999). The impact of connectives and anaphoric expressions on expository discourse comprehension. *Document Design*, *1*, 39-51.

Delin, J., Hartley, A., & Scott, D. (1996). Towards a contrastive pragmatics: Syntactic choice in English and French instructions. *Language Sciences, 18*(3-4), 897-931.

Delin, J., Bateman, J., & Allen, P. (2002). A model of genre in document layout. *Information Design Journal, 11*(1), 54-66.

Di Eugenio, B. (1998). An action representation formalism to interpret natural language instructions, *Computational Intelligence*, 14(1), 89-133.

Firth, J.R, (1969). *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press. (firth edition: 1957).

Foley, W. A., & Van Valin Jr., R. D. (1985). Information packaging in the clause. In T. Shopen (Ed.), *Syntactic Typology and Linguistic Description 1*, (pp. 282-364), Cambridge: Cambridge University Press.

Fries, P. H. (1995). Patterns of information in initial position in English. In P. H. Fries & M. Gregory (Eds.), *Discourse in society: Systemic functional perspectives. Meaning and choice in language: Studies for Michael Halliday* (pp. 47-66). Norwood: Ablex.

Garcia-Debanc, C. (Ed.) (2001). Les discours procéduraux, *Langages 141* (mars 2001).

Garcia-Debanc, C. & Grandaty, M. (2001). Incidence des variations de la mise en forme textuelle sur la compréhension et la mémorisation de textes procéduraux (règles de jeux) par des enfants de 8 à 12 ans. *Langages 141*, 92-104.

Goldman, A. I. (1970). *A theory of human action*. Englewood Cliffs, N.J.: Prentice Hall.

Gomez-Gonzàlez, M.-A. (2001). *The theme-topic interface - Evidence from English*. Amsterdam: John Benjamins.

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

Grote, B. (1999). Decomposing discourse relations for instructional texts. In *Proceedings of Levels of Representation in Discourse Workshop*, University of Edinburgh (http://www.hcrc.ed.ac.uk/~lorid99).

Gruber, T. R. (1991).The role of common ontology in achieving sharable, reusable knowledge bases. *In Proceedings of the 2nd International Conference on the Principles of Knowledge Representation and Reasoning*. (pp. 601-602).

Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5, 2. 199-220.

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

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. (pp. 539-545).

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics* 30.135-175.

Lambrecht, K. (1994). *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Luc, C., Mojahid, M., Virbel, J., Garcia-Debanc, C., & Péry-Woodley, M.-P. (1999). A linguistic approach to some parameters of layout: A study of enumerations. In R. Power

& D. Scott (Eds.). In *AAAI 1999 Fall Symposia: Using Layout for the Generation, Understanding or Retrieval of Documents* (pp. 20-29). North Falmouth, Massachussetts.

Luc, C., Mojahid, M., Péry-Woodley, M.-P., & Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. *In* M. Gaio & E. Trupin (Eds.), Proceedings of CIDE 2000 (Colloque International sur le Document Électronique), (21-40). Lyon, France. Europia Productions.

Luc, C., & Virbel, J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. *Verbum, 23*(1), 103-123.

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Mani, I. (2001). *Automatic Summarization*. Amsterdam: John Benjamins.

Mann, W. C. & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243-281.

Mann, W.C. & Thompson, S.A. (Eds.) (1992). *Discourse description. Diverse linguistic analyses of a fund-raising text*. Amsterdam: John Benjamins.

Marcu. (2001). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.

Meyer, I. (2000). Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework ». In D. Bourigault, C. Jacquemin, & M.-Cl. L'homme (Eds.), *Recent advances in computational terminology* (pp. 279-302). Amsterdam: John Benjamins.

Minel, J.-L. (2003). *Filtrage sémantique. Du résumé automatique à la fouille de textes*. Paris: Hermès-Lavoisier.

Nunberg, G. (1990). *The linguistics of punctuation.* Menlo Park: Center for the Study of Language and Information.

Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins.

Péry-Woodley, M.-P. (2001). Modes d'organisation et de signalisation dans des textes procéduraux. *Langages 141*, 28-46.

Péry-Woodley, M.-P. (2005). Organisation discursive des textes procéduraux : caractériser des segments naturels pour un accès sélectif. In D. Alamargot, P. Terrier, & J.-M. Cellier (Eds.), *Production, compréhension et usage des écrits techniques au travail* (pp. 31-47). Toulouse: Octarès Éditions.

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

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.) *Radical pragmatics* (pp. 223-255). New-York: Academic Press.

Redeker, G. (1991). Linguistic markers of discourse structure. *Linguistics*, *29*, 1139-1172.

Rey, A. (1995). *Essays on terminology*. Amsterdam: John Benjamins.

Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam: John Benjamins.

Sanders, T., & Noordman, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes, 29*(1), 37-60.

Sanders, T. J. M., & Spooren, W. (2001). Text representation as an interface between language and its users. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 1-28). Amsterdam: John Benjamins.

Scott, D., Delin, J., & Hartley, A. (1998). Identifying congruent pragmatic relations in procedural texts. *Languages in Contrast, 1*(1), 45-82.

Skuce, D., & Meyer, I. (1991). Terminology and knowledge engineering: Exploring a symbiotic relationship. In *Proceeedingss of the 6th International Worshop on Knowledge Acquisition for Knowledge-Based Systems*. (29/1-29/2).Banff.

Swales, J.-M. (1990). *Genre analysis, English in academic and research settings*. Cambridge: Cambridge University Press.

Thompson, S. A. (1985). Grammar and written discourse: Initial vs. final purpose clauses in English. *Text, 5*(1-2), 55-84.

Todorov, T. (1984). *Mikhail Bakhtin: the dialogical principle*, (translated by Wlad Godzich). Minneapolis: University of Minnesota Press.

Torrance, M., & Bouayad-Agha , N. (2001). Rhetorical structure analysis as a method for understanding writing processes. In *Proceedings of 4th International Multidisciplinary Approaches to Discourse (MAD) Workshop*. (51-59).Ittre, Belgium.

Townsend, D. J. (1997). Processing clauses and their relationships during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing Interclausal Relationships - Studies in the Production and Comprehension of Text* (pp. 265-282). London: Lawrence Erlbaum Associates.

Trosborg, A. (2000). *Analysing professional genres*. Amsterdam: John Benjamins.

Vander Linden, K. & Martin, J.H. (1995). Expressing rhetorical relations in instructional text: a case study of the purpose relation. *Computational Linguistics*, 21 (1), 29-58.

Virbel, J. (1989). The contribution of linguistic knowledge to the interpretation of text structures. In J. André, V. Quint & R. K. Furuta (Eds), *Structured Documents* (pp. 161-181). Cambridge: Cambridge University Press.

Virtanen, T. (1992a). *Discourse functions of adverbial placement in English*. Åbo: Åbo Akademi University Press.

Virtanen, T. (1992b). Given and new information in adverbials: Clause initial adverbials of time and place. *Journal of Pragmatics, 17*(2), 99-117.

Walker, M., Joshi, A. & Prince, E. (1998) (Eds.). *Centering Theory in Discourse*. Oxford: Clarendon Press.

---

[i] This distinction between descriptive and theoretical studies is clearly an over-simplification, however, as some descriptive studies of technical documents have the potential to lead to the elaboration or revision of models of text organisation (e.g. Thompson 1985).

[ii] Examples from Gomez-González (2001), with capitals indicating stress. Note that in some languages, such as French, this contrast is likely to be realised syntactically.

[iii] See also Di Eugenio (1998) for a formal computational approach to the representation of actions in instructional texts.

[iv] Much information can also be obtained from the RST website: http://www.sfu.ca/rst/index.html.