Representing discourse for automatic text summarization via shallow NLP techniques

Laura Alonso i Alemany

Departament de Lingüística General
Universitat de Barcelona

Barcelona, 2005

Advisors

Irene Castellón Masalles
Universitat de Barcelona

Lluís Padró Cirera
Universitat Politècnica de Catalunya

there is life beyond perfection

Abstract

In this thesis I have addressed the problem of text summarization from a linguistic perspective. After reviewing some work in the area, I have found that many **satisfactory approaches to text summarization rely on general properties of language that are reflected in the surface realization of texts**.

The main claim of this thesis is that some general properties of the discursive organization of texts can be identified at surface level, that they provide objective evidence to support theories about the organization of texts, and that they can be of help to improve existing approaches to text summarization.

In order to provide support for this claim, I have determined which **shallow cues are indicative of discourse organization**, and, of these, which fall under the scope of the natural language processing techniques that are currently available for Catalan and Spanish: punctuation, some syntactical structures and, most of all, discourse markers.

I have developed a framework to obtain a **representation of the discursive organization of texts of inter- and intra-sentential scope**. I have described the nature of minimal discourse units (discourse segments and discourse markers) and the relations between them, relating theoretical explanations with empirical descriptions, systematizing discursive effects as signalled by shallow cues.

Based on the evidence provided by shallow cues, I have **induced an inventory of basic meanings to describe the discursive function of relations between units as signalled by shallow cues**. The inventory is organized in two dimensions of discursive meaning: structural (*continuation* and *elaboration*) and semantic (*revision, cause, equality* and *context*). In turn, each dimension is organized in a range of markedness, with a default meaning that is assigned to unmarked cases.

I have shown that the proposed representation **contributes to improve the quality of automatic summaries** in two different approaches: it was integrated within a lexical chain summarizer, to obtain a representation of text that combines cohesive and coherence aspects of texts, and it was also one of the components of the analysis of text in an e-mail summarizer.

Finally, some experiments with human judges indicate that this representation of texts is useful to **explain how some discursive features influence the perception of relevance**, more concretely, to characterize those fragments of texts that judges tend to consider irrelevant. Experiments with automatic procedures for the analysis of texts correlate well with the perception of relevance observed in human judges.

Acknowledgements

If a thesis supposes both a personal progress and a scientific contribution, I feel the balance between the two is far more advantageous for me than for the scientific community, as will be reflected in these acknowledgements.

In the first place, I have to say that this work would have never been possible without the financial support of the Spanish Research Department, which paid for four years of my learning under grant PB98-1226.

First of all, a special acknowledgement goes to the people who have revised and read part or the whole of this thesis, specially my advisors and the programme committee, because they had to suffer this ability that runs in my family, that we make the telling of a story last longer than the story itself.

I want to sincerely thank are my supervisors, Irene Castellón and Lluís Padró. Their support never failed, despite difficulties at all levels. They proved to be a neverending source of sense, perspective and encouragement. I wish I can some time be able to express my gratitude for all their help.

During this time I have worked with many other people who helped me come closer to being a scientist. Karina Gibert helped me to put some order in the chaos I had in front of me, I learny to have fun with achievements with Bernardino Casas, Gemma Boleda had and kept the will to continue working until we could reach something good, and I learnt so much from and with Maria Fuentes that I could not possibly write it here. Moreover, a lot of people annotated texts just because it is good to help the others, specially Ezequiel and Robert, who spent many sunday afternoons thinking about causes and consequences.

An important part of the fact that this thesis was begun, pursued and (finally!) finished has to be attributed to my attending academic events, where people made me feel part of a community and made my work worth, and where I could have a clearer idea of what I could expect from research and what I wanted to avoid. Of these, the three wise men provided most valuable advice in the plainest way. Horacio Rodríguez kept reminding me that the main purpose of NLP is that we have jobs, but never failed to recognise the beauty of a good explanation, the interest of unrealizable ideas and the use of (some of) our implementations. Toni Badia was most kind in helping me to interpret things that were far beyond my perspective. Henk Zeevat made me feel that not knowing things was not necessarily a shame, but a circumstance, and that getting to know things required some effort but was worthwhile, and never found it difficult to spend as much time as necessry talking, until things made sense. Finally, the memory of my grandfather assured me that it was important to try to carry out solid work, regardless of how far I might get.

Most specially, I would like to thank all those people who, without having anything to do from my work, suffered directly from the fact that I was working. First of all, my dear Òscar, who suffered my reading on Sundays, my stress for deadlines and all my anxiety and uncertainties, but enjoyed almost nothing of my maturity. Then, Maria had to listen to ununderstandable talk about my research, renouncing for years to the wonderful apple-talk we have finally recovered. Josep had to listen to my constant mhms and ouchs and brrs, but we always managed to share precious silences. My family hardly remembered my face when I began visiting them again, but my brother and my mother always provided support in the best way they knew: loads of home-made food, piles of references to apparently useless technicalities, and most of all a strong basis for my tired and weak principles and unconditional love.

Finally, I thank Gabriel, for furnishing my feet with solid ground, so that we could walk on to the last stages of this work.

Contents

Chapter 1

Introduction

Automatic text summarization (AS) deals with the process of reducing one or more texts to a summary text that condensates the most relevant information given a certain information need, by concatenating literal fragments of the orginal text(s) or by identifying key concepts and generating a new text, all this by means of automatic techniques for natural language processing (NLP).

1.1 Motivation

The automatic summarization of texts is an area of research that has attracted the interest of many researchers through the years, because it can contribute to gain a better understanding of the way people produce and understand language, because it can solve the growing needs for information synthesis in our society or just because "*it is difficult*" (Spärck-Jones 2004). Probably this last reason can explain the interest of the area better than any other: AS supposes a challenge for the capabilities of current NLP and it addresses unsolved questions about basic linguistic and semiotic aspects of texts and about the faculty of human language in general.

From a purely applied perspective, AS is one of the applications of NLP that aim to act as an interface between the particular information needs of particular persons and the huge amount of information that is publicly available to satisfy this information, specially on the World Wide Web. Many techniques have been developed to address this problem: information retrieval, question answering, information condensation, information synthesis, and, of course, text summarization. It is often difficult to establish a clear distinction between these different applications, and they tend to merge: text summarization has been addressed as a kind of complex question answering (DUC 2005) or it has been used as an aid for retrieval of structured documents (Kan 2003).

From a theoretical point of view, AS can be considered as an experimental method to investigate the adequacy of theoretical proposals of the way people produce and understand texts. If we assume that people summarize texts based on a certain representation of their organization (Kintsch and van Dijk 1983; Sanders and Spooren 2001), the performance of a systematic procedure as AS upon such representation can contribute to validate hypotheses about text organization. Indeed, some linguistic theories of text organization

(Grosz and Sidner 1986; Mann and Thompson 1988; Polanyi 1988) have been applied to represent texts as the basis for AS systems to produce summaries.

Despite the many efforts devoted to it, the problem of how to summarize texts is still far to be solved. On the one hand, it seems that the capabilities of NLP systems are insufficient to provide summaries of texts that are comparable to summaries produced by humans. This goal seems to require understanding and generation capabilities that are well beyond what can be achieved with the current state-of-the-art tools and resources.

On the other hand, the whole process of summarization is still unclear. The computational process of summarizing texts has been divided in two major steps: analyzing the input text(s) and producing a summary, but it is not clear how any of these two steps should be addressed. One of the main reasons for this is the fact that the quality of different summaries for a given text cannot be properly evaluated. The task of summarization seems to be so intrinsically interpretative that different people usually produce very different summaries for a given text, and judge the quality of automatic summaries also very differently. As a consequence, it cannot be assessed whether a given technique introduces an improvement to the process of summarization, because it cannot be assessed whether the summaries produced by this technique are better than without it.

However, the situation is not desperate. The basic target of most current summarization techniques consists in identifying units of content in a source text, and finding which are more relevant within the text or to satisfy a given information need; these will constitute the summary. Some techniques are clearly helpful to determine the relevance of content units, for example, techniques based on the frequency of words or in the position of units within a structure of the text, or those exploiting the presence of certain cue words. In contrast, the contribution of more sophisticated techniques, like rhetorical or template-driven analysis still remains to be assessed.

In this thesis I address the problem of AS from a linguistic perspective. I focus on the analysis of the input text(s) as a determining factor to improve text summarization. My starting hypotheses are:

1. A representation of texts at discourse level is useful for AS systems to improve the quality of resulting summaries.

2. Such representation can be built based on evidence found at the surface realization of texts.

   (a) Systematic relations can be established between this evidence and the behaviour of human judges to summarize texts.

   (b) This evidence can be the basis to obtain (part of) the targeted representation of discourse by shallow NLP techniques, more concretely, by those techniques available for Catalan and Spanish.

In Figure 1.1 we can see an illustration of how a certain discursive representation of texts exploiting surface properties can be useful to condensate texts. In this example, a

[$_1$ En este caso,] [$_2$ [$_3$ y] [$_4$ gracias al] excelente trabajo de la antropóloga Silvia Ventosa,] [$_5$ autora de "Trabajo y vida de las corseteras de Barcelona",] [$_6$ esta leyenda urbana se comprobó] [$_7$ que era un calco de una historia] [$_8$ que conmocionó a la localidad francesa de Orleans] [$_9$ en 1969] .

In this case, and thanks to the excellent work of the anthropologist Silvia Ventosa, author of "Work and life of Barcelona's seamstresses", this urban legend was found to be a copy of a story that shook the French town of Orleans in 1969.

summary:
esta leyenda urbana se comprobó que era un calco de una historia
this urban legend was found to be a copy of a story

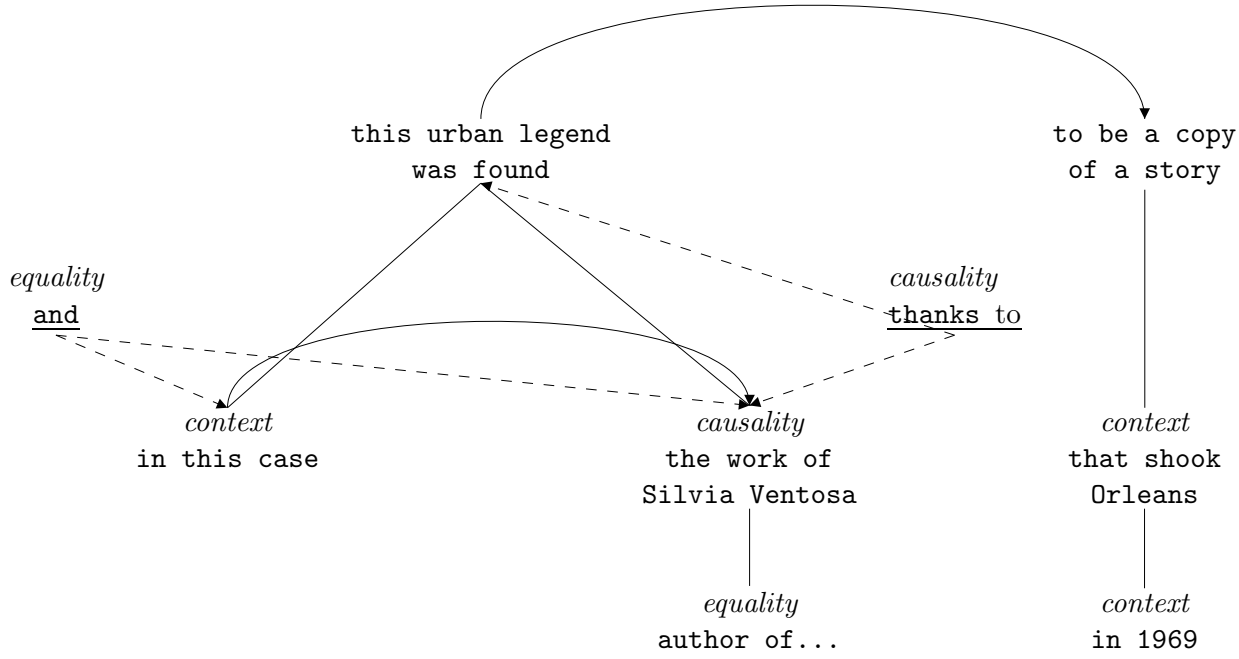


Figure 1.1: Example of how a text can be summarized by taking the most relevant content units as determined by its discursive structure.

complex sentence has been represented as a local discourse structure, where elements in the structure represent units of content, hierarchical structure represents the relations between topics and subtopics in the text and tags in italics represent the coherence relation between two given units, elicited by discourse markers like *and* or *thanks to*. This structure can be taken as indicative of the relevance of the content units; in the example, the topmost units in the hierarchical structure of the text are taken as the most relevant and thus constitute a "summary" for the sentence.

However, this is not the only possible summary, others could also be considered good summaries, for example:

> `esta leyenda urbana se comprobó que era un calco de una historia que conmocionó a la localidad francesa de Orleans en 1969`

or

> `gracias al excelente trabajo de la antropóloga Silvia Ventosa, esta leyenda urbana se comprobó que era un calco de una historia`

This example serves to illustrate that the assignment of relevance is unclear, even if it is based on a representation of discourse. However, we claim that such representation should be useful to model the processes of human summarization better, and to make decisions of automatic systems more principled.

1.2 Aim of the thesis

The final goal of this thesis is to provide a representation of discourse that improves text summarization and can be obtained by shallow NLP. This general goal can be divided in the following research desiderata:

1. Propose an overall organization of texts to capture those aspects of their discursive organization that are useful for text summarization, elicit implicit assumptions about coherence and relevance, and provide a finer-grained, deeper insight on the organization of text beyond clausal level.

   (a) Take into account key concepts from previous work.

   (b) Identifiy and typify (minimal) units at discourse level.

   (c) Determine which properties of these units provide information that is useful for AS, focussing specially in their coherence relations, because this aspect of discourse is very unclear in current work.

   (d) Determine the relative relevance of these units according to their role in the organization of discourse that we are analyzing.

2. Assess the role of linguistic mechanisms at discourse level for text summarization (both for producing summaries and for evaluating them), strongly based on empirical data, more concretely, on the surface realization of texts.

   (a) Determine surface textual cues that are indicative of discourse organization, focussing in coherence relations.

   (b) Determine the scope to which the information provided by these clues is reliable.

   (c) Induce an inventory of discourse relations from these shallow cues.

   (d) Characterize discourse units by shallow cues.

3. Provide an approach to summarization that can be obtained with knowledge-poor NLP resources, but at the same time provides a representation of texts that can be naturally integrated with knowledge-rich approaches, thus contributing to build the gap between shallow and deep NLP.

   (a) Determine which textual cues can be exploited by shallow techniques to obtain the targeted representation of discourse with a reliable degree of certainty.

   (b) Adapt the representation of discourse proposed theoretically to meet the limitations of shallow techniques.

4. Develop methods and resources to apply this approach to AS in Catalan and Spanish, and also to English.

   (a) Create resources storing the shallow cues found useful for AS via shallow NLP, and the information associated to them.

   (b) Develop methods to identify and characterize new units using data mining methods.

   (c) Develop algorithms for identification and characterization of units and relations between them, exploiting the above resources and the systematicities found in texts.

   (d) Identify systematic relations between the organization of discourse obtained by shallow techniques and the way people summarize texts.

   (e) Systematize these relations algorithmically, to exploit discourse organization for AS.

5. Apply the obtained representation of discourse as an aid to AS systems, and assess its contribution to improve the quality of the resulting summaries.

1.3 Summary of the thesis

In Chapter 2 I make an analysis of the problem of text summarization. First I discuss evaluation of summaries, a major problem in the area. I describe various evaluation methods, and discuss the achievements of most recent methods based on the comparison of automatic summaries with an inventory of content units obtained from human summaries. The main advantages of these methods with respect to previous evaluations are that they are based on content overlap vs. string overlap, that they are based on multiple summaries vs. a single gold standard and that they achieve an important degree of agreement between different human evaluators. I propose some improvements to these methods, namely, to complete the inventory of content units to which automatic summaries are compared by identifying minimal content units in the source texts and assigning a certain relevance score to them. To do that, an automated discourse analysis as the one I propose in this thesis can be exploited.

After a brief overview of different approaches to AS, I argue that approaches that aim at understanding, using deep processing, are costly, typically non-robust, with small coverage, and very often encounter conflicts between different kinds of knowledge. In contrast, robust approaches are based on shallow processing of texts, but they exploit properties of texts in an unprincipled way, which makes theoretical implications difficult. Moreover, it is arguable that these shallow approaches have an upper bound, although this is difficult to assess because of the evaluation of summaries is a problem still to be solved.

Intuitively, one would think that deeper analyses should improve text summarization. However, it seems that very simple techniques, relying on shallow features of texts, like the distribution of lexic or the presence of some cue words are capable of producing summaries of a quality comparable to approaches that use much deeper knowledge. One possible explanation for this phenomenon is that shallow and deep analyses of text capture the same kind of linguistic phenomena, and thus the representation of text that they yield is very comparable. This would explain why summaries would not significantly differ.

I argue that shallow approaches need not be limited, but they can be very useful to gain deeper understanding if they are explicitly and systematically related to a theory of the organization of texts. For example, in the analysis of discursive structure by means of shallow cues (Marcu 1997b; Schilder 2002; Reitter 2003a), the evidence is gathered by shallow techniques, but the representation obtained is perfectly coherent with what would have been obtained with much more complex methods and resources, although probably not as rich.

The discursive aspect of texts seems to provide an adequate representation of texts to be summarized. To support this claim, I assess the contribution of a very shallow representation of discourse organization to the improvement of automatic summaries for Spanish.

Chapter 3 specifies the representation of discourse by which texts will be described via shallow NLP techniques to improve text summarization. I define the structure by which discourse is represented, and I the basic discourse units: segments and markers. To define

these three concepts I take into account previous computational work.

I specify a structure of discourse adapted to the limitations of shallow NLP, in that rich representations are obtained wherever there is enough evidence to do that, and default relations are established between units or constructs when there is no reliable evidence to obtain richer structures. Thus, the presence of shallow cues is crucial to obtain rich representations, and the scope to which the information provided by these cues is reliable determines the scope of rich relations, and leaves the rest to be accounted for by default relations.

As a result, discourse is represented as a sequence of content-rich hierarchical trees of local scope, related with each other by default relations. Moreover, we establish that this representation can be multidimensional, so that heterogenous meanings are captured by independent structures, establishing different relations between the same set of discourse units. Multidimensionality allows to capture some configurations of discourse relations that escape traditional tree-like representations of discourse. However, in the implementation via shallow NLP that we propose in this thesis multidimensionality is not necessary, because such configurations of discourse cannot be captured.

As for basic discourse units, discourse segments are basic units of content, while discourse markers convey procedural information about the organization of content units in a structure.

In Chapter 4, I address the crucial question of determining an inventory of meanings to be used as labels to describe the semantics of discourse relations. In the first place, I determine which particular shallow cues are indicative of discourse relations, and especially of the semantics of those relations: punctuation, some syntactical structures and, most of all, discourse markers. These cues will be taken as the empirical evidence to determine the nature of discourse relations.

I determine that the information provided by these cues is reliable only at a short scope, more concretely, at inter-sentential and intra-sentential. This is the scope of the local structures within which we can identify content-rich relations; these local structures will be related to each other by default, content-poor discourse relations.

Basically following Knott (1996), the meaning of relations is divided in distinct features, so that the global meaning of a relation is constituted compositionally as a conglomerate of such relations. I propose that such features are organized in dimensions grouping heterogeneous meanings, so that a given feature can co-exist with features in other dimensions, but not in its same dimension. Each of these dimensions is in turn organized in a range of markedness, with the interesting property that the least marked meaning is taken by default to characterize unmarked cases, where no shallow cue can be found that provides any information about the organization of discourse in that part of the text.

Also following Knott (1996), I have induced a small set of such features based on the evidence provided by discourse markers. The main difference with Knott is that I only exploit highly grammaticalized discourse markers, because my aim is to establish basic, uncontroversial areas where discursive meaning is clearly distinct. Thus, I only provide a gross-grained description of the semantics of discourse relations. These features have been used to characterize a small lexicon of prototypical discourse markers in Catalan, Spanish

and English, presented in Appendix A.

In Chapter 5 empirical evidence is provided to support the theoretical proposal of the two previous chapters. I present two experiments with human judges and two experiments with automatic procedures.

In the first experiment, 35 human judges had to summarize texts by removing intrasentential fragments, ranging from words to whole clauses. We show that, for this task, humans do not agree much more than could be expected by chance. However, when we model the texts by the discourse organization presented before, we obtain that in some cases, judges show a very important ratio of agreement, up to the point that the task can be said to be reproducible. We obtain that the model that explains best the behaviour of humans, in terms of inter-judge agreement, is a combination between the semantics of the relation that a given discourse segment holds with other segments, and some surface features of its context of occurrence, like the occurrence of punctuation signs.

In the second experiment, three judges identify relations between minimal discourse units, and they label them by their semantics. A significant degree of agreement between judges is achieved, even despite the fact that two of them are naive. We believe that this experiment supports the claim that the proposed meanings are basic, because they can capture the intuitions of naive judges.

Then, we show that discourse segments can be reliably identified by automatic procedures, with varying degrees of precision and recall, depending on the kind of information exploited. We find that discourse markers are the most informative and reliable information to identify segments. In the last experiment we show that previously unseen discourse markers can be automatically identified by lexical acquisition techniques, exploiting some properties that characterize them in large amounts of text.

Finally, in Chapter 6 I conclude this thesis by summarizing its contributions and sketching lines of research that are left for future work.

Note: all examples of English in this thesis have been taken from the British National Corpus, if the contrary is not indicated. Examples of Catalan and Spanish have been taken from the electronic version of the newspaper *El Periodico de Catalunya*. Translation of the examples can be found in the separata accompanying this thesis.

Chapter 2

Approaches to text summarization

In this chapter we present our field of research, automatic text summarization (AS). We analyze the current state of the question, and highlight two major issues of concern to us: the fact that the evaluation of summaries is still an unsolved problem, and the fact that solid summarization systems have to be based in robust, general and principled methods that capture basic insights of the organization of language at textual level.

Then, we propose an approach to AS that commits to these requirementes, based on an analysis of the discursive level of texts. In order to be robust, this analysis is limited to what can be obtained by shallow NLP techniques, more concretely, to the capacities of NLP for Catalan and Spanish. Since it is aimed to be insightful, general and principled, it is based on a general model about the organization of discourse, developed in the rest of the thesis.

The structure of the chapter is as follows. In the first place, in Section 2.1 we discuss the interest of automatic text summarization as a research field, and we analyze the problem from a computational perspective. We discuss how, despite many clarification efforts, automatic text summarization seems to be still more ill-defined than many other tasks in NLP.

In Section 2.2 we address the problem of evaluating the quality of different summaries for the same text(s), which is currently a burning issue in text summarization, because the achievements of different systems cannot be properly assessed, and thus working efforts cannot be properly directed. We make a critique of the methods that are currently considered as most reliable for evaluation of summaries, totally manual and based on the analysis of multiple human summaries for the same text(s). We argue that these methods could benefit from an analysis of the source text that identifies its basic content units and associates them to a relevance score based on their linguistic form.

Section 2.3 is an overview of different approaches to automatic text summarization over the last fifty years. We do not aim to be exhaustive, but we analyze those approaches that introduce different perspectives into the problem, contributing to a better understanding. We provide a taxonomy of the main approaches to summarization, with examples of current summarization systems, and trying to analyze the achievements and limitations of each of these approaches. We conclude this review of work in the area trying to synthesize what

makes a good approach to summarization.

In Section 2.4 we argue that the analysis of the discursive organization of texts is compliant with the properties that make an approach to summarization successful. We show that a certain representation of discourse can be successfully integrated within robust, insightful approaches to text summarization, improving the final results. We exemplify this in two AS applications: an integration of the proposed analysis of discourse and lexical chains to obtain general-purpose summaries, and an e-mail summarizer one of whose modules provides an analysis of discourse structure that significantly contributes to the quality of the final summary.

Finally, in Section 2.5 we describe our general framework to obtain an analysis of discourse structure for text summarization via shallow NLP techniques.

2.1 Introduction: factors in automatic text summarization

In the last years there has been a growing interest in automatic text summarization (AS), because it is expected that it can provide useful solutions to growing needs in our society, like synthesizing huge amounts of information so as to diminish the amount of processing left for humans to do, locating information relevant to a concrete information need, discriminating relevant and irrelevant information and, in general, discovering information which may be relevant.

Besides this applied interest, there are more fundamental research interests in AS. The process whereby people summarize texts seems to involve core human cognitive abilities, because in order to summarize a text, a person seems to have both its informational content and its textual and documental properties available (Kintsch and van Dijk 1983; Endres-Niggemeyer 1998), which requires complex understanding processes. It can be expected that AS systems are based on a model of the way people look for information, understand it, integrate it to fulfill their information needs, and eventually produce a synthetic (linguistic) representation of what they consider relevant information. If so, AS systems can be seen as devices to support or refutate theoretical claims about crucial aspects of the way people understand texts.

From the point of view of Artificial Intelligence, AS is a complex, high-level task, so it supposes a challenge by itself, with the added interest of an immediate practical application. More concretely, within NLP it requires the integration of many existing NLP tools and resources in an emulation of human behaviour. The optimal solution should take advantage of all necessary resources at hand, but not more than necessary, it should be as close as possible to the nature of the object and it should address as closely as possible the intended objective.

2.1.1 Text summarization from a computational perspective

There have been many efforts devoted to analyze the problem of AS and systematize the process (Gladwin *et al.* 1991; Endres-Niggemeyer *et al.* 1993; Sparck-Jones 1993; Sparck-Jones 1997; Hovy and Marcu 1998; Sparck-Jones 1999b; Mani and Maybury 1999; Radev 2000; Hahn and Mani 2000; Sparck-Jones 2001a; Hovy 2001; Baldwin *et al.* 2000; Mani 2001), here we will describe the aspects that most have been generally considered essential for a good understanding of the problem.

From a computational perspective, "*A summary is a reductive transformation of a source text into a summary text by extraction or generation*" (Sparck-Jones 2001b).

The problem of summarization has traditionally been decomposed into three phases:

- *analyzing* the input text to obtain text representation,
- *transforming* it into a summary representation,
- and *synthesizing* an appropriate output form to generate the summary text.

Much of the early research on AS has been devoted to the analysis of the source text, probably for two main reasons: first, a summary that fulfills an information need can be built by simple concatenation of literal fragments of the source text. Second, the generation of natural language expressions automatically requires huge NLP resources, which are far beyond the capabilities of most of the current research groups in the area even today. As it is, the analysis of the source text can currently be considered as the main factor influencing the quality of the resulting summary. Indeed, once the crucial aspects of texts have been identified and characterized, virtually all heuristics for selection of relevant information will perform well.

Besides the internal steps for building a summary, many contextual factors affect the process of summarization, mostly concerning the kind and number of documents to be summarized, the medium of communication, the expected format of the summary, the intended audience, etc. Effective summarising requires an explicit and detailed analysis of context factors, since summaries are configured by the information need they have to fulfill. Sparck-Jones (1999a) distinguishes three main aspects of summaries: input, purpose and output, which we develop in what follows.

2.1.1.1 Input Aspects

The features of the text to be summarized crucially determine the way a summary can be obtained. The following aspects of input are relevant to the task of TS:

Document Structure Besides textual content, heterogeneous documental information can be found in a source document, for example, labels that mark headers, chapters, sections, lists, tables, etc. If it is well systematized and exploited, this information can be of use to analyze the document. For example, Kan (2002) (Kan 2003) exploits the organization of medical articles in sections to build a tree-like representation of the source. Teufel and Moens (2002) (Teufel and Moens 2002b) systematize the

structural properties of scientific articles to assess the contribution of each textual segment to the article, in order to build a summary from that enriched perspective.

However, it can also be the case that the information it provides is not the target of the analysis. In this case, document structure has to be removed in order to isolate the textual component of the document.

Domain Domain-sensitive systems are only capable of obtaining summaries of texts that belong to a pre-determined domain, with varying degrees of portability. The restriction to a certain domain is usually compensated by the fact that specialized systems can apply knowledge intensive techniques which are only feasible in controlled domains, as is the case of the multidocument summarizer SUMMONS (McKeown and Radev 1995), specialized in summaries in terrorism domain applying complex Information Extraction techniques. In contrast, general purpose systems are not dependant on information about domains, which usually results in a more shallow approach to the analysis of the input documents.

Nevertheless, some general purpose systems are prepared to exploit domain specific information. For example, the meta summarizer developed at Columbia University (Barzilay *et al.* 1999; Barzilay *et al.* 2001; Hatzivassiloglou *et al.* 1999; Hatzivassiloglou *et al.* 2001; McKeown *et al.* 2002) applies different summarizers for different kinds of documents: MULTIGEN (Barzilay *et al.* 1999; McKeown *et al.* 1999) is specialized in simple events, DEMS (Schiffman *et al.* 2001) (with the bio configuration) deals with biographies, and for the rest of documents, DEMS has a default configuration that can be resorted to.

Specialization level A text may be broadly characterized as ordinary, specialized, or restricted, in relation to the presumed subject knowledge of the source text readers. This aspect can be considered the same as the *domain* aspect discussed above.

Restriction on the language The language of the input can be general language or restricted to a sublanguage within a domain, purpose or audience. It may be necessary to preserve the sublanguage in the summary.

Scale Different summarizing strategies have to be adopted to handle different text lengths. Indeed, the analysis of the input text can be performed at different granularities, for example, in determining meaning units. In the case of news articles, sentences or even clauses are usually considered the minimal meaning units, whereas for longer documents, like reports or books, paragraphs seem a more adequate unit of meaning. Also the techniques for segmenting the input text in these meaning units differ: for shorter texts, orthography and syntax, even discourse boundaries (Marcu 1997a) indicate significant boundaries, for longer texts, topic segmentation (Kozima 1993; Hearst 1994) is more usual.

Media Although the main focus of summarization is textual summarization, summaries of non-textual documents, like videos, meeting records, images or tables have also been

undertaken in recent years. The complexity of multimedia summarization has prevented the development of wide coverage systems, which means that most summarization systems that can handle multimedia information are limited to specific domains or textual genres (Hauptmann and Witbrock 1997; Maybury and Merlino 1997). However, research efforts also consider the integration of information of different media (Benitez and Chang 2002), which allow a wider coverage of multimedia summarization systems by exploiting different kinds of documental information collaboratively, like metadata associated to video records (Wactlar 2001).

Genre Some systems exploit typical genre-determined characteristics of texts, such as the pyramidal organization of newspaper articles, or the argumentative development of a scientific article. Some summarizers are independent of the type of document to be summarized, while others are specialized on some type of documents: healthcare reports (Elhadad and McKeown 2001), medical articles (Kan 2003), agency news (McKeown and Radev 1995), broadcast fragments (Hauptmann and Witbrock 1997), meeting recordings (Zechner 2001), e-mails (Muresan *et al.* 2001; Alonso *et al.* 2003a), web pages (Radev *et al.* 2001), etc.

Unit The input to the summarization process can be a *single document* or *multiple documents*, either simple text or multimedia information such as imagery audio, or video (Sundaram 2002).

Language Systems can be language-independant, exploiting characteristics of documents that hold cross-linguistically (Radev *et al.* 2003; Pardo *et al.* 2003), or else their architecture can be determined by the features of a concrete language. This means that some adaptations must be carried out in the system to deal with different languages. As an additional improvement, some multi-document systems are able to deal simultaneously with documents in different languages (Chen 2002; Chen *et al.* 2003).

2.1.1.2 Purpose Aspects

Situation TS systems can perform general summarization or else they can be embedded in larger systems, as an intermediate step for another NLP task, like Machine Translation, Information Retrieval or Question Answering. As the field evolves, more and more efforts are devoted to task-driven summarization, in detriment of a more general approach to TS. This is due to the fact that underspecification of the information needs supposes a major problem for design and evaluation of the systems. As will be discussed in Section 2.4.2.2.1, evaluation is a major problem in TS. Task-driven summarization presents the advantage that systems can be evaluated with respect to the improvement they introduce in the final task they are applied to.

Audience In case a user profile is accessible, summaries can be adapted to the needs of specific users, for example, the user's prior knowledge on a determined subject. *Background* summaries assume that the reader's prior knowledge is poor, and so

extensive information is supplied, while *just-the-news* are those kind of summaries conveying only the newest information on an already known subject. Briefings are a particular case of the latter, since they collect representative information from a set of related documents.

Usage Summaries can be sensitive to determined uses: retrieving source text (Kan *et al.* 2001), previewing a text (Leuski *et al.* 2003), refreshing the memory of an already read text, sorting...

2.1.1.3 Output Aspects

Content A summary may try to represent all relevant features of a source text or it may focus on some specific ones, which can be determined by queries, subjects, etc. *Generic* summaries are text-driven, while *user-focused* (or query-driven) ones rely on a specification of the user's information need, like a question or key words.

Related to the kind of content that is to be extracted, different computational approaches are applied. The two basic approaches are top-down, using information extraction techniques, and bottom-up, more similar to information retrieval procedures. Top-down is used in query-driven summaries, when criteria of interest are encoded as a search specification, and this specification is used by the system to filter or analyze text portions. The strategies applied in this approach are similar to those of Question Answering. On the other hand, bottom-up is used in text-driven summaries, when generic importance metrics are encoded as strategies, which are then applied over a representation of the whole text.

Format The output of a summarization system can be plain text, or else it can be formatted. Formatting can be targeted to many purposes: conforming to a pre-determined style (tags, organization in fields), improving readability (division in sections, highlighting), etc.

Style A summary can be *informative*, if it covers the topics in the source text; *indicative*, if it provides a brief survey of the topics addressed in the original; *aggregative*, if it supplies information non present in the source text that completes some of its information or elicits some hidden information (Teufel and Moens 2002b); or *critical*, if it provides an additional valoration of the summarized text.

Production Process The resulting summary text can be an *extract*, if it is composed by literal fragments of text, or an *abstract*, if it is generated. The type of summary output desired can be relatively polished, for example, if text is well-formed and connected, or else more fragmentary in nature (e.g., a list of key words).

There are intermediate options, mostly concerning the nature of the fragments that compose extracts, which can range from topic-like passages, paragraph or multiparagraph long, to clauses or even phrases. In addition, some approaches perform editing operations in the summary, overcoming the incoherence and redundancy often found

in extracts, but at the same time avoiding the high cost of a NL generation system. Jing and McKeown (2000) (Jing and McKeown 2000) apply six re-writing strategies to improve the general quality of an extract-based summary by edition operations like deletion, completion or substitution of clausal constituents.

Surrogation Summaries can stand in place of the source as a surrogate, or they can be linked to the source (Kan *et al.* 2001; Leuski *et al.* 2003), or even be presented in the context of the source (e.g., by highlighting source text, (Lehmam and Bouvet 2001)).

Length The targeted length of the summary crucially affects the informativeness of the final result. This length can be determined by a compression rate, that is to say, a ratio of the summary length with respect to the length of the original text. Traditionally, compression rates range from 1% to 30%, with 10% as a preferred rate for article summarization. In the case of multidocument summarization though, length cannot be determined as a ratio to the original text(s), so the summary always conforms to a pre-determined length. Summary length can also be determined by the physical context where the summary is to be displayed. For example, in the case of delivery of news of summaries to hand-helds (Boguraev *et al.* 2001; Buyukkokten *et al.* 2001; Corston-Oliver 2001), the size of the screen imposes severe restrictions to the length of the summary. Headline generation is another application where the length of summaries is clearly determined (Witbrock and Mittal 1999; Daumé III *et al.* 2002). In very short summaries, coherence is usually sacrificed to informativeness, so lists of words are considered acceptable (Kraaij *et al.* 2002; Zajic *et al.* 2002).

2.2 Evaluation of summaries: a wall to be climbed

In any NLP task, evaluation is crucial to assess the achievements of different approaches, and so to assess which lines of research are worth pursuing because they lead to improvements in performance. Indeed, in tasks like information retrieval or machine translation, evaluation is a crucial aspect of progress in the field.

In text summarization, evaluation is still to be solved, and so the whole area suffers from a lack of general assessment to direct research, since it is unclear which approaches provide good results.

The first large-scale initiative for evaluation of automatic summaries was carried out in 1998, with the SUMMAC contest. Different AS systems submitted monodocument summaries for the same set of texts. The summaries were evaluated *extrinsically* and *intrinsically*. While intrinsic evaluation focussed in summaries by themselves, extrinsic evaluation assessed how well the summary could subrogate the original text, by assessing how well human judges could answer a set of relevant questions about a text, comparing performance with the original text, with a human summary and with an automatic summary.

The results from these two evaluations were confusing, but what became clear was that evaluation of summaries was a non-trivial task and that much effort had to be devoted to it.

Extrinsic evaluation was soon discarded as a valid method for evaluation, because the bias introduced by the final application may obscure insights about the nature of summaries. Besides, extrinsic evaluation is very costly. For these reasons in the large-scale effort for evaluation of summarization systems within the framework of the DUC contest of summarization (`http://nist.gov/duc`), evaluation of summaries was carried out intrinsically, applying a standard *gold standard* approach, where results from automatic systems were compared to summaries produced by humans using the SEE environment

But intrinsic evaluation is not free of problems. Indeed, it was found that there was much variability on the ratings provided by different judges for a same summary, regardless of automatic or human. It was seen that general-purpose summarization is by itself an ill-defined task, because the usual scenario is that there are many good summaries for a given text. So, the usual gold standard approach is not applicable, because there is more than one possible correct summary to be compared with.

Since this problem was clear, efforts have focussed in capturing the goodness of automatic summaries taking into account many possible gold standards. Within the DUC contest, many summaries were produced for each text or set of texts to be summarized, and each summary was compared to one of them at random. The method of comparison is based in the SEE system (Summary Evaluation Environment, Lin 2001), and consists in dividing both human and automatic summaries into minimal units, establishing equivalences in meaning between units in the automatic and the human summary, and determining the coverage of the meaning in the human summary that is achieved by the automatic summary. It has to be noted that the units of human and automatic summaries are of different granularities: while the first are elementary discourse units (EDUs, Carlson, Marcu and Okurowski 2003), automatic summaries are splitted into sentences, which may contain more than one EDU.

Even despite this highly delimited evaluation process, Lin and Hovy (2002a) showed that there was big variability between the judgements produced by human judges about the same summary, which makes the assessments of quality based on human judges unreliable. In contrast, Lin and Hovy (2003) also showed that string overlap correlated with human judgements to assess similarity between human and automatic summaries, and created a tool to evaluate automatic summaries by n-gram overlap with human summaries, ROUGE (Lin 2004), based in the idea implemented in BLEU for the evaluation of automatic translations (Papineni, Roukos, Ward and Zhu 2001).

Criticisms have been made to ROUGE because it is based in string overlap, which arguably fails to capture relations in meaning without a correspondence in form, like entailment, inclusion, synonymy, etc. Moreover, it is capable of making differences between clearly bad and clearly good summaries, but it cannot make finer-grained differences, so it often happens that the results of many systems are rated as of undistinguishable quality.

We tried to overcome this limitation in Alonso, Fuentes, Rodríguez and Massot (2004d), where we assessed the quality of automatic summaries for the DUC 2003 corpus by extrap-

olation from the scores assigned to the n most similar summaries for the same text, which had been manually assigned by NIST assessors. Similarity between automatic summaries and model summaries was based in unigram overlap, but arguably this method is able to capture more similarities in meaning than ROUGE for two main reasons: because the amount of summaries that can be compared is bigger (we can compare with all the automatic summaries produced for DUC 2003) and because the summaries that are compared are all automatic, most of them are extraction-based, and so they are formed by parts of the source text. Since the source text provides a limited amount of source forms, the similarities in meaning of extractive summaries are expected to be more strongly correlated to similarities in form than human summaries, which can contain totally different forms to convey a same meaning.

In the last years, two new approaches to evaluation of summaries are emerging, that proposed by van Halteren and Teufel (2003) and that of Nenkova and Passonneau (2004). Both are based in the comparison between content units that can be found in human summaries that are used as a gold standard, but there are two main differences with the approach in SEE. First, the information overlap is not determined as a proportion of the unit, but as a full match between the content in a unit in the automatic summary and in the human summaries, which was the main source of disagreement between judges using SEE. Second, multiple human summaries are taken into account to build the inventory of units to compare automatic summaries with.

Evaluation methods proposed recently are promising in that they try to capture the goodness of automatic summaries by comparison not to one, but to many human summaries. van Halteren and Teufel (2003) and Nenkova and Passonneau (2004) try to identify basic units of content (*factoids* and *summary content units, SCUs*, respectively) by comparison of different human summaries for a same text. These basic conent units are assigned a degree of relevance relative to their frequency of occurrence in the set of human summaries. Then, summaries are evaluated by identifying their units of content, mapping these units to the set of units established by comparison of human summaries, and then assigning them the corresponding relevance score. The overall relevance of a summary is a sum of the relevance of all its units.

These methods present many advantages. First, they are content-based, a method of comparison arguably superior to string-based comparison. Then, they try to capture the vagueness inherent to summarization by taking into account not only one, but many summaries. The gold standard against which summaries are compared is formed by all the units of content that can be found in the human summaries that are taken into account. Moreover, the agreement between judges is much higher in these methods that for the evaluation based in SEE, probably because these establish one-to-one mappings between units in the gold standard and units in the summary to be evaluated, while in SEE the relation is a percentage.

Despite these interesting properties, we see two main shortcomings to these methods: that they are totally manual, thus costly and prone to subjectivity, and that they are fully based on summaries, totally neglecting the features of the source text. This last issue seems to be at the root of one of the main problems of these methods: that the inventory of

content units that can be induced from various summaries is not exhaustive, that is, that previoulsy unseen content units can be found in new summaries, worryingly, in summaries to be evaluated.

van Halteren and Teufel (2004) state that the number of content units in the gold standard formed by the content units found in all the summaries for a given text grows as the number of summaries increases, and it doesn't seem to come to a stable point even if as many as 10 or 20 summaries are taken into account. Thus, it is probable that a summary to be evaluated contains units to which no relevance score can be assigned.

An alternative to these methods was presented by Amigó *et al.* (2004). They present an empirical study of the similarity between "good" summaries, produced by humans, and "bad" summaries, produced automatically. They found that the best measure to distinguish these two kinds of summaries is indeed based on content overlap, instead of word overlap, but, in contrast with the methods presented so far, it envisages a method to obtain these concepts automatically, thus drastically reducing the cost of evaluation, and reducing also the possibility of infinite increase in the number of content units to be compared. For these reasons, this method is taken as a reference for evaluation in the next DUC contests.

However, this method still has the drawback of relying entirely on human-produced abstracts, and therefore it cannot properly evaluate the goodness of previously unseen content units.

We argue that a basic discursive analysis of the source text, identifying minimal discourse units and some of their relations, could improve the assessment of the goodness of a given summary with respect to the source text. This additional source of information could contribute to minimize the two major problems pointed out above.

In the first place, a basic discursive analysis that relies on shallow textual cues can be automated with a good degree of reliability, thus making the task of identifying content units and their relations lighter and arguably more systematic, although also more error-prone.

Moreover, a basic discursive analysis is useful to identify more content units than those occurring in the summaries. This kind of analysis is also capable of associating very rough relevance scores to some of these units, if there is shallow evidence to do it. Interestingly, it is irrelevant units that are more marked with shallow evidence, so this kind of analysis would provide precisely the evidence complementary to that provided by the human-summary-based gold standard: a set of content units associated to low relevance. Thus, this kind of analysis could be very useful to identify and characterize *bad* summaries, which, in a context where new techniques are tried out, is as useful as identifying good summaries.

2.3 State of the art in automatic text summarization

In this Section, we will present previous work in the field of automated text summarization, a very active field of research which experienced an important growth in the late 1990s

and beginning of 2000s. That's why many comparative studies can be found in the literature (Paice 1990; Zechner 1997; Sparck-Jones 1999a; Hovy and Marcu 1998; Tucker 1999; Radev 2000; Maybury and Mani 2001). Also the SUMMAC and DUC contests provide a good overview of the most innovative working systems over the years.

We will only discuss here those works that made a qualitative contribution to the progress of the field. There are several ways in which one can characterize different approaches to text summarization. In Alonso *et al.* (2003c) we presented a comparison between three possible classifications of text summarization systems (Mani and Maybury 1999; Tucker 1999 and Alonso 2001), and classified a number of summarization systems according to each of them.

Here we will provide a brief overview of the classification presented in (Alonso 2001), updated with recent systems. This classification is based in the kind of information exploited to build the summary, and will be useful to illustrate our point that successful approaches to summarization are precisely those that exploit general linguistic mechanisms. Three main kinds of approaches are distinguished: those exploiting lexical aspects of texts, those working with structural information and those trying to achieve deep understanding of texts. However, most current systems are not limited to a single feature of the text, but produce summaries by combining heterogeneous kinds of information, as discussed in Section 2.3.4.

2.3.1 Lexical information

These approaches exploit the information associated to words in the texts. Some of them are very shallow, relying on the frequency of words, but some others apply lexical resources to obtain a deeper representation of texts. Beginning by the most shallow, the following main trends can be distinguished.

Word Frequency approaches assume that the most frequent words in text are the most representative of its content, and consequently fragments of text containing them are more relevant. Most systems apply some kind of filter to leave out of consideration those words that are very frequent but not indicative, for example, by the *tf*idf* metric or by excluding the so-called *stop words*, words with grammatical but no meaning content. The oldest known AS system exploited this technique (Luhn 1958b), but it is also the basis of many current commercial systems, like those integrated in text processing applications, for example.

Domain Frequency tries to determine the relevance of words by first assigning the document to a particular domain. Domain specific words have a previous relevance score, which serves as a comparison ground to adequately evaluate their frequency in a given text. This approach is not highly wide spread, since the process of determining topics is costly and involves important decisions on the design of the system, as shown by its best known example, SUMMARIST (Lin 1998).

Concept Frequency abstracts from mere word-counting to concept-counting. By use of an electronic thesaurus or WordNet, each word in the text is associated to a more general concept, and frequency is computed on concepts instead of particular words.

Cue words and phrases can be considered as indicators of relative relevance or non-relevance of fragments of text in respect to the others. This approach has never been used as the basis of a system, but many systems consider the information of cue words to improve the relevance score assigned to units, as Edmunson (1969) originally did.

Chains can be built from lexical items which are related by conceptual similarity according to a lexical resource (*lexical chains*) or by identity, if they co-refer to the same entity (*co-reference chains*). The fragments of text crossed by most chains or by most important chains or by most important parts of chains can be considered the most representative of the text, as shown in the classic approach of Barzilay (1997).

A common assumption of these approaches is that repeated information is a good indicator of importance.

2.3.2 Structural information

A second direction in AS tries to exploit information from the texts as structured entities. Since texts are structured in different dimensions (documental, discursive, conceptual), different kinds of structural information can be exploited. Beginning by the most shallow:

Documental Structure exploits the information that texts carry in their format, for example, headings, sections, etc. This approach has resulted extremely useful for highly structured domains, like medicine articles, as shown in Kan (2003), where a whole structure of topics can be induced based mostly on the information provided by section headings and lexical repetition in subdomains of medicine. Also Teufel and Moens (2002a) exploit this information, among other kinds of information, to obtain a rhetorical analysis of scientific articles in the area of computational linguistics.

Textual Structure Some positions in text systematically contain the most relevant information, for example, the beginning paragraph of news stories. These positions are usually genre- or domain-dependent, but automatic procedures can be exploited to identify them, as shown in (Lin and H.Hovy 1997).

Conceptual structure The chains mentioned in lexical approaches can be considered as a kind of conceptual structure.

Discursive Structure can be divided in two main lines: linear or narrative and hierarchical or rhetoric. The first tries to account for *satisfaction-precedence*-like relations among pieces of text, the second explains texts as trees where fragments of text are

related with each other by virtue of a set of rhetorical relations, mostly asymmetric, as used for text summarization by Ono, Sumita and Miike (1994a) and Marcu (1997b).

2.3.3 Deep understanding

Some approaches try to achieve understanding of the text in order to build a summary. Two main lines can be distinguished:

Top-down approaches try to recognize pre-defined knowledge structures to texts, for example, templates or frames, as is the case of the system SUMMONS (McKeown and Radev 1995), which tries to fill MUC-like templates, and some multidocument summarization systems, which try to fill templates for pre-defined kinds of events, like MULTIGEN (Barzilay *et al.* 1999; McKeown *et al.* 1999), specialized in simple events, DEMS (Schiffman *et al.* 2001), specialized in biographies, or GLEANS (Marcu *et al.* 2002), which has different configurations for different kinds of events.

Bottom-up approaches try to represent texts as highly conceptual constructs, such as scene. Others apply fragmentary knowledge-structures to clue parts of text (Kan and McKeown 1999) and then build a complete representation out of these small parts.

2.3.4 Approaches combining heterogeneous information

The approaches that seem to produce the best summaries according to DUC evaluation are those that combine heterogenous sources of information to obtain a representation of the text, including virtually all of the techniques above that are within the NLP capabilities of each research group. In some cases, these approaches integrate deep representations of texts, obtained by performing knowledge-rich analyses, but in many other cases a combination of simple, knowledge-poor analyses (Lin and Hovy 2002b) seems to produce results of a quality comparable to that of knowledge-rich systems, thus presenting a rather interesting ratio effort–results.

This approach has many interesting properties. For example, most of the systems carry out a study of the contribution of different kinds of knowledge to the quality of the final summaries. Many also take advantage of machine learning techniques to (partially) determine the best weighting of the different features. Moreover, integrating many different aspects of documents, even if one of them is privileged above the others, guarantees a complete representation of the source text, in accordance with its intrinsic complexity.

The first approach to integrate heterogeneous information to obtain summaries is Edmunson (1969), who enhances the work of Luhn (1958a) by incorporating equally simple techniques. It integrates naturally the textual dimension with the documental one, also handles different linguistic levels. He envisages a more complex treatment of the object (the text), to achieve an improved objective (the summary), but he does not increase the complexity of the strategy, as was usual in the moment, by creating a single, more complex model, but by integrating in a natural manner different simple models. This is a particular

case of a strategy that has proved successful in AS in general, because of its simplicity, flexibility and because it seems able to capture insightful properties of text.

In several systems (Kupiec *et al.* 1995; Teufel and Moens 2002b; Hovy and Lin 1999; Mani and Bloedorn 1999) different summarization methods are combined to determine the relevance of units: title-based relevance scoring, cue phrases, position in the text, and word frequency.

As the field progresses, summarization systems tend to use more and deeper knowledge, but the tendency to exploit heterogeneous information to determine the relevance of units collaboratively still remains. So, the tendency is that heterogeneous kinds of knowledge are merged in increasingly enriched representations of the source text(s).

These enriched representations allow for adaptability of the final summary to new summarization challenges, such as multidocument, multilingual and even multimedia summarization. In addition, such a rich representation of text is a step forward generation or, at least, pseudo-generation by combining fragments of the original text. Good examples of this are (McKeown *et al.* 2002; Lin and Hovy 2002c; Daumé III *et al.* 2002; Lal and Rueger 2002; Harabagiu and Lacatusu 2002), among others.

2.3.5 Critical overview of summarization techniques

No strong conclusions can be drawn from this overview because, as explained in the previous section, the standard evaluation methods up to the last big scale evaluation effort, in DUC 2004, were uncapable of making the fine-grained distinctions that are needed to analyze the contribution of different techniques to summary quality. We are hopeful that recent proposals for evaluation by content overlap will make it possible to evaluate systems in DUC 2005.

We will consider that the usage of a summarization technique accross systems is indicative of its utility for producing high-quality summaries. We can see that some techniques, even if they are very simple (word frequency, presence of cue words), have been exploited in a big number of systems, while others that looked very promising a priori (template filling, entity-driven information extraction) have been applied in very rare occasions.

Techniques that have been applied most often are usually very simple, we can consider that the most widespread approaches are those that exploit shallow textual correlates of relevance, like distribution of words in texts (frequencies, lexical chains, etc.), location in relevant positions (titles, beginning or end of texts, beginning of paragraphs) or the presence of certain cue words indicating relevance or lack of relevance. The main limitation of these approaches is that they cannot obtain a deep representation of text, but only shallow indications of the relevance of particular textual units.

Approaches exploiting shallow evidence as a mere indicator of relevance do not contribute to significant progress in the area because they do not increase our understanding of the problem and do not provide an insightful representation of source texts, even if they may produce improvements in particular summaries. However, as we discuss below, these limitations can be overcome by relating shallow evidence to a general theory of text organization, which allows to obtain richer representations of text by accumulation of various

kinds of shallow evidence.

In contrast, many of the techniques that have not lived longer than as a prototype aim to obtain a deep representation of the text, as is the case of approaches based in information extraction techniques to fill a pre-defined template. Information extraction approaches have the advantage that texts can be represented as a scene, where one can clearly identify relevant entities, their attributes and their relations. Their main disadvantages are that all knowledge has to be pre-defined, which represents a very important effort of development and, more worryingly, a substantial lack of flexibility, since previously unseen cases cannot be treated. Thus, these approaches suffer from a restricted coverage and a lack of robustness. In contrast with the limitations of shallow techniques, these limitations are rooted in the essence of the method, and cannot be overcome.

One last critical remark on previous work on text summarization concerns the identification of content units. As said before, the analysis of the source text is a key step to improve the selection of content to make a summary. Most approaches represent texts as a set of content units associated to a relevance score and. In some cases, relations between units are also identified, indicating relevance but also of coherence relations between units. Determining which units to identify is then a crucial aspect of summarization, but most approaches are focussed in the assignment of relevance and neglect the aspect of determining what might be a unit.

Thus, we can conclude that a number of approaches have been applied to achieve relevant and readable text summaries, but that it is difficult to assess their contribution to the quality of the final summaries because evaluation of summaries is an issue still to be solved. In any case, it seems clear that satisfactory approaches to AS exploit general properties of documents to guarantee robustness of the strategy, while leaving room to gain deeper insight into the configuration of texts. We have found that shallow approaches are interesting because they seem to provide results of a quality comparable to that of approaches exploiting deep knowledge techniques, while having wider coberture, and they are more flexible to treat unseen cases.

Altough it is true that shallow techniques can only provide shallow representations of the source text, it is also true that they are progressively gaining in depth, and it can be expected that they provide much more insight on the structure of language in the future.

Under our point of view, the progress of shallow techniques is crucially subject to the principledness by which they are worked out. If evidence of relevance is collected and not organized, the kind of information one can obtain from it will always be the same. In contrast, if we relate shallow cues with a theory of the organization of texts, the information they provide can be increasingly enriched, because it can be integrated with other evidence, so that they can collaboratively configure progressively richer representations of the source text. Lack of principledness affects also the characterization of content units in the representation of the source text.

As a conclusion, it seems clear that a sensible approach to summarization consists in exploiting shallow evidence in a principled way. In the following section we argue that shallow evidence related to a theory of discourse is compliant with these requirements.

2.4 Discourse for text summarization

As follows from what we have exposed until now, satisfactory approaches to AS exploit general properties of texts. We argue that a discursive representation can be specially useful, for two main reasons: because the units of content for summarization seem to be best characterized at discourse level, more than at clausal level, and because the relevance of content units depends on their configuration within the representation of the text as a stand-alone structure, if no other factor is provided to assess relevance (like for example query).

Those aspects of a discursive representation of text that we consider useful for summarization are:

- identify **content units** at discursive level. In the case of an extractive summarizer, they are the basic unit of which summaries are built, and they are may also be useful for summary evaluation, as discussed in Section 2.2.
- determine the relative **relevance** of each content unit with respect to its local context and, eventually, to the whole document. Relevance assessment contributes to determine which units contain the information that is to be conveyed in the summary.
- identify **coherence** relations between units. These relations are of use to direct or constrain the selection of units to obtain a coherent summary. As we have explained in Section 2.4.2.2, in some contexts summaries are required to be strongly coherent.

Relevance and coherence are crucial aspects of a representation of text for AS. The diverse approaches to summarization rely on definitions of these concepts directly determined by the kind of information that they can obtain from a text: lexical items, presence of Named Entities, presence of cue phrases, co-reference chains, position of elements in the document, etc. For example, Salton, Singhal, Mitra and Buckley (1997) define relevance as directly proportional to the number of nodes departing from a certain meaning unit in a graph where nodes are units and arcs are lexical relations between the lexical items contained in discourse units.

A general definition of these concepts is provided by Mani (2001):

> *Salience (or relevance) is the weight attached to information in a document, reflecting both the document content as well as the relevance of the document information to the application. [...]*
>
> *Coherence [is] the way the parts of the text gather together to form an integrated whole.*
>
> *(Mani 2001, pg. 11)*

In a structural representation as the one we propose in Chapter 3, the relative relevance of a unit is relative to its structural context of occurrence. For example, in a tree-like structure like that proposed by the Rhetorical Structure Theory (RST, Mann and Thompson,

described below), proximity to the root is indicative of relevance. Also units that are highly related with other units are more relevant, according to general principles of graph theory as applied to summarization by Salton, Singhal, Mitra and Buckley (1997). In case discourse structure has labelled relations, signalling the semantics of the relations between discourse constituents, the label of the relations can also factor out the final relevance count for each segment.

As for **Coherence**, and the closely related concept of **cohesion**, they has often been considered as one of the basic features that constitute text. Halliday and Hasan (1976), one of the basic references for the formal treatment of texts, define cohesion as *the set of possibilities that exist in the language for making text hang together* (Halliday and Hasan 1976). Coherence has often been described by resorting to world knowledge and complex reasoning mechanisms, but also with respect to the textual mechanisms that realize it in the linguistic surface form.

For AS, coherence relations enhance the process of content selection, providing for the felicity or discursive well-formedness of the resulting summary text. Given a selection of units based on relevance, coherence relations specify which units are required to be included in the summary content, regardless of their own relevance.

2.4.1 Previous work in exploiting discourse for text summarization

There have been various approaches to exploit the discursive organization of text to improve the relevance and quality of final summaries. Many approaches exploit discursive properties in an unprincipled way, for example, by removing all subordinated clauses, by including the sentence immediately preceding a sentence introduced by a discourse marker, etc. As we have said, we feel that these approaches do not contribute to significant progress in the area because they do not increase our understanding of the problem and do not provide an insightful representation of source texts, even if they may produce improvements in particular summaries.

There have been quite some approaches basing summaries on a representation of the discursive aspect of texts. Some of these approaches exploit deep understanding of texts, others are based on shallow evidence. We will specially focus in the latter, because, as we have said, shallow evidence seems more adequate to address the task of text summarization, and it is also closer to our own approach. The most popular theory of text organization underlying summarization approaches has been the RST.

Ono, Sumita and Miike (1994b) are the first known researchers to have applied RST to analyze a text as a hierarchical tree, where the relevance of discourse units is relative to their proximity to the root of the tree. They propose that a summary of the text can be obtained exploiting one of the properties of the *nuclearity* principle of rhetorical relations *à la* RST: the fact that, in an asymmetric relation between two discourse units, where one, the *nucleus*, is more important than the other, the *satellite*, the least important unit can be removed, preserving the main aim of the text. However, their proposal has not been

pursued further on.

Some approaches produce summaries exploiting rich representations of discourse that are based on deep analyses of text produced by the NLP tools owned by private corporation. This is the case of Corston-Oliver (1998) at Microsoft and Polanyi *et al.* (2004) at Xerox.

Corston-Oliver (1998) applies an RST approach to represent the structure of text, and applies this representation to summarization. The rhetorical analysis of texts is based in the deep analysis provided by Microsoft language processing tools, which is said to reach the level of propositional analysis and even allows to establish inferential relations between clauses.

Polanyi *et al.* (2004) produce summaries applying a set of heuristics to a representation of discourse based on the Linguistic Discourse Model (Polanyi 1988; Polanyi 1996). Just as Corston-Oliver (1998), this analysis is based on a structure of discourse that builds directly upon sentential syntax and semantics, and contemplates different kinds of discourse structuring devices, such as basic hierarchical and linear structuring, genre-determined schemata and interactional frameworks. It relies on the language processing tools of Xerox Parc.

Another interesting approach to summarization based on deep analyses of the discursive structure of text is that of Hahn (1990). Text is represented as in a hierarchy of thematic units obtained by analyzing it with a knowledge base of the domain. Then, summaries can be obtained by retrieving different granularities of this structure. This approach is clearly domain-dependent, since it strongly relies on the existence of a knowledge base to obtain the hierarchical representation of the thematic units of the text. A similar approach is proposed by Kan (2003), with the main difference that Kan also provides a methodology to induce the knowledge base from the texts themselves.

Some other approaches rely on shallow cues to obtain a discursive representation of text, these are closer to what we consider an approach useful to introduce progress in the field of text summarization and are also closer to our own approach, so we will discuss them deeper.

Marcu (1997b) is the best known application of the RST-based approach to summarization, because he provides a thorough description of the procedures he applies and also of the resources he exploits. He implements an RST-based discourse parser for English that makes no use of world knowledge, but instead is fully based on shallow textual evidence, namely, discourse markers and word form co-occurrence. The aim of this parser is to obtain a discourse structure that conforms to the well-formedness requirements of RST (Mann and Thompson 1988).

Marcu's parser is totally text-based, it does not depend on domain-dependant sources of knowledge, but exploits the general properties of discourse markers. This guarantees robustness of the system, but, in contrast to other shallow approaches, partial analyses are not allowed in case a complete one can not be provided.

The architecture of the parser leaves room for incorporating heterogeneous information on discourse structure; for example, word co-occurrence is used to identify cohesion-based relations between discourse units. However, the fact that the approach is relation-based makes it difficult to incorporate heterogeneous discourse information, because any new

information must be expressed in terms of relations and the new relations must be comparable to the previous ones. This makes it difficult, for example, to incorporate such useful information for discourse processing as information structure, topics, etc. Also the constraints imposed by RST itself have been criticised for lack of descriptive adequacy, as will be argued in the next chapter.

Soricut and Marcu (2003) follows the philosophy of Marcu (1997b) but applies a machine learning approach to build the discourse parser. A corpus with syntactic and RST-based rhetorical annotation (Carlson, Marcu and Okurowski 2001) serves as the basis for the learning process. The performance of the machine learning parser is better than for the manually built one, probably because rules are based on the objective weighting of a number of examples, instead of the subjective impression of the analyst, and also because the machine learning parser is able to exploit much more information on the examples, more concretely, the lexico-syntactic structure of the training examples.

Schilder (2002) implements an underspecified version of SDRT (Lascarides and Asher 1993) to obtain a representation of discourse that he argues can be useful to summarize texts. A two-step strategy combining deep and shallow approaches is applied. First, rich structures are obtained with a hand-crafted, grammar-based analysis, but only for those parts of text where discursive clues are found. Then, less informative structures, but covering the whole text, are derived with a more robust strategy based on word forms.

More concretely, discourse particles provide information to apply the rules of an underspecified version of SDRT (Lascarides and Asher 1993) that determines inmediate dominance, dominance, precedence and equivalence relations between minimal discourse constituents, obtaining rhetorical schemata of *local* scope.

Higher-level discourse structure should be treated with more complex knowledge (intentions, beliefs, plans, genres, etc.). Since neither this knowledge nor the ability to deal with it are available, Schilder obtains higher order relations between discourse units via a topicality measure, an adaptation of the $tf * idf$ metric. Then, the relations between these schemata are constrained by a topicality measure based on the *tf*idf* index to determine the relative relevance of each schema.

In essence, Schilder's approach exploits the same kind of information as Marcu (1997b), in short: rich structure is derived from available discourse markers and word-based measures account for the relations between those units with no discourse markers. However, Schilder's and Marcu's parsers crucially differ in their ability to integrate heterogeneous discourse knowledge. The stepwise analysis of Schilder (2002) allows the progressive enrichment of the resulting structure, while keeping relative independence between the information to be taken into account.

Finally, we would like to mention an interesting approach to discourse analysis, although it has not yet been applied to summarization. Its interest lies in the fact that it establishes a very principled connection between shallow evidence and an insightful theoretical framework. DLTAG (Forbes, Miltsakaki, Prasad, Sarkar, Joshi and Webber 2003) incorporates the syntactic and semantic properties of discourse markers upon sentential analysis to go beyond sentential level and reach what they call *low level discourse structure and discourse semantics*.

The function of discourse markers in structure derivation is to establish relations between textual entities of various sizes (ranging from clauses to the whole previous text) and kind (from syntax-based units to abstract objects (Asher 1993)). Some discourse markers, like *although* or *because*, take their arguments structurally, and some others anaphorically, like *however* or *in that case*. For the second kind, it is problematic to determine the referential argument with precision.

Within DLTAG, no specific machinery for the analysis of discourse is developed, but rather the existing mechanisms at clausal level are adapted for the discursive level. This supposes an important economy of development that allows high portability of the system. However, as in the case of the two parsers presented so far, the coverage of the system is critically limited to the amount of discourse markers that have been accounted for.

In sum, many approaches have exploited an analysis of discourse in AS systems, to improve the quality of the resulting systems. Some of these approaches are based on shallow textual clues. Discourse markers are among the most used of these clues, because they are highly informative of the relations between discourse units. Notwithstanding, other kinds of information, like topicality measures or lexico-syntactic structures improve the accuracy of the analysis.

2.4.2 Assessing the utility of shallow discourse analysis for text summarization

In this section we will describe two AS systems that incorporate the discourse analysis that will be presented in the following chapters. The description of the performance of this systems will try to illustrate in detail how a certain representation of the discursive structure of texts, even if it is fully based in shallow evidence, can significantly improve the quality of the resulting summaries.

2.4.2.1 Integrating cohesion and coherence for text summarization

In Alonso and Fuentes (2002) and Alonso and Fuentes (2003) we carried out a series of experiments integrating *cohesive* properties of text with *coherence* relations, to obtain an adequate representation of text for automatic summarization. More concretely, a summarizer based on lexical chains (Fuentes and Rodríguez 2002) was enchanced with rhetorical and argumentative structure obtained via discourse markers.

2.4.2.1.1 Previous Work on Combining Cohesion and Coherence Traditionally, two main components have been distinguished in the discursive structure of a text: cohesion and coherence. As defined by Halliday and Hasan (1976), **cohesion** tries to account for relationships among the elements of a text. Four broad categories of cohesion are identified: *reference*, *ellipsis*, *conjunction*, and *lexical cohesion*. On the other hand, **coherence** is represented in terms of relations between text segments, such as *elaboration*, *cause* or *explanation*. Mani (2001) argues that an integration of these two kinds of discursive information would yield significant improvements in the task of text summarization.

Corston-Oliver and Dolan (1999) showed that eliminating discursive satellites as defined by the Rhetorical Structure Theory (RST) Mann and Thompson (1988), yields an improvement in the task of Information Retrieval. Precision is improved because only words in discursively relevant text locations are taken into account as indexing terms, while traditional methods treat texts as unstructured bags of words.

Some analogous experiments have been carried out in the area of TS. Brunn *et al.* (2001) and Alonso and Fuentes (2002) claim that the performance of summarizers based on lexical chains can be improved by ignoring possible chain members if they occur in irrelevant locations such as subordinate clauses, and therefore only consider chain candidates in main clauses. However, syntactical subordination does not always map discursive relevance. For example, in clauses expressing finality or dominated by a verb of cognition, like *Y said that X*, the syntactically subordinate clause *X* is discursively nuclear, while the main clause is less relevant (Verhagen 2001).

In Alonso and Fuentes (2002), we showed that identifying and removing discursively motivated satellites yields an improvement in the task of text summarization. Nevertheless, we will show that a more adequate representation of the source text can be obtained by ranking chain members in accordance to their position in the discourse structure, instead of simply eliminating them.

2.4.2.1.2 Summarizing with Lexical Chains The lexical chain summarizer follows the work of Morris and Hirst (1991) and Barzilay (1997).

As can be seen in Figure 2.1 (left) the text is first segmented, at different granularity levels (paragraph, sentence, clause) depending on the application. To detect chain candidates, the text is morphologically analysed, and the lemma and POS of each word are obtained. Then, Named Entities are identified and classified in a gazzetteer. For Spanish, a simplified version of Palomar *et al.* (2001) extracts co-referenece links for some types of pronouns, dropping off the constraints and rules involving syntactic information.

Semantic tagging of common nouns is been performed with *is-a* relations by attaching EuroWordNet (Vossen 1998) synsets to them. Named Entities are been semantically tagged with *instance* relations by a set of *trigger words*, like *former president*, *queen*, etc., associated to each of them in a gazzetteer. Semantic relations between common nouns and Named Entities can be established via the EWN synset of the trigger words associated to a each entity.

Chain candidates are common nouns, Named Entities, definite noun phrases and pronouns, with no word sense disambiguation. For each chain candidate, three kinds of relations are considered, as defined by Barzilay (1997):

- **Extra-strong** between repetitions of a word.
- **Strong** between two words connected by a direct EuroWordNet relation.
- **Medium-strong** if the path length between the EuroWordNet synsets of the words is longer than one.

Being based on general resources and principles, the system is highly parametrisable. It has a relative independence because it may obtain summaries for texts in any language for which there is a version of WordNet an tools for POS tagging and Named Entity recognition and classification. It can also be parametrised for obtaining summaries of various lengths and at granularity levels.

As for relevance assessment, some constraints can be set on chain building, like determining the maximum distance between WN synsets of chain candidates for building medium-strong chains, or the type of chain merging when using gazetteer information. Once lexical chains are built, they are scored according to a number of heuristics that consider characteristics such as their length, the kind of relation between their words and the point of text where they start. Textual Units (TUs) are ranked according to the number and type of chains crossing them, and the TUs which are ranked highest are extracted as a summary. This ranking of TUs can be parametrised so that a TU can be assigned a different relative scoring if it is crossed by a strong chain, by a Named Entity Chain or by a co-reference chain. For a better adaptation to textual genres, heuristics schemata can be applied.

However, linguistic structure is not taken into account for scoring the relevance lexical chains or TUs, since the relevance of chain elements is calculated irrespective of other discourse information. Consequently, the strength of lexical chains is exclusively based on lexic. This partial representation can be even misleading to discover the relevant elements of a text. For example, a Named Entity that is nominally conveying a piece of news in a document can present a very tight pattern of occurrence, without being actually relevant to the aim of the text. The same applies to other linguistic structures, such as recurring parallelisms, examples or adjuncts. Nevertheless, the relative relevance of these elements is usually marked structurally, either by sentential or discursive syntax.

2.4.2.1.3 Incorporating Rhetorical and Argumentative Relations The lexical chain summarizer was enhanced with discourse structural information as can be seen in Figure 2.1 (right).

Following the approach of Marcu (1997b), a partial representation of discourse structre was obtained by means of the information associated to a discourse marker lexicon. Discourse markers are described in four dimensions:

- **matter**: three different kinds of subject-matter meaning are distinguished, namely *causality*, *parallelism* and *context*.
- **argumentation**: three argumentative moves are distinguished: *continuation*, *elaboration* and *revision*.
- **structure**: following the notion of right frontier Webber (1988, Polanyi (1988), *symmetric* and *asymmetric* relations are distinguished.
- **syntax**: describes the relation of the discourse marker with the rest of the elements at the discourse level

The information stored in this discourse marker lexicon was used for identifying inter- and intra-sentential discourse segments (Alonso and Castellón 2001) and the discursive relations holding between them. Discourse segments were taken as Textual Units by the Lexical Chain summarizer, thus allowing a finer granularity level than sentences.

Two combinations of discourse marker descriptive features were used, in order to account for the interaction of different structural information with the lexical information of lexical chains. On the one hand, *nucleus-satellite* relations were identified by the combination of *matter* and *structure* dimensions of discourse markers. This **rhetorical** information yielded a hierarchical structure of text, so that satellites are subordinate to nucleus and they are accordingly considered less relevant. On the other hand, the **argumentative** line of text was traced via the *argumentation* and also *structure* discourse marker dimensions, so that segments were tagged with their contribution to the continuation of the argumentation.

These two kinds of structural analyses are complementary. Rhetorical information is mainly effective at discovering local coherence structures, but it is unreliable when analyzing macro-structure. As Knott *et al.* (2001) argue, a different kind of analysis is needed to track coherence throughout a whole text; in their case the alternative information used is focus, we have opted for argumentative orientation. Argumentative information accounts for a higher-level structure, although it doesn't provide much detail about it.

2.4.2.1.4 Experiments A number of experiments were carried out in order to test whether taking into account the structural status of the textual unit where a chain member occurs can improve the relevance assessment of lexical chains (see Figure 2.2). Since the discourse marker lexicon and the evaluation corpus were available only for Spanish, the experiments were limited to that language. Linguistic pre-processing was performed with the CLiC-TALP system (Atserias *et al.* 1998a; Arévalo *et al.* 2002).

For the evaluation of the different experiments, the evaluation software MEADeval (MEAD) was used, to compare the obtained summaries with a golden standard. From this package, the usual precision and recall measures were selected, as well as the simple cosine. Simple cosine (simply *cosine* from now on) was chosen because it provides a measure of similarity between the golden standard and the obtained extracts, overcoming the limitations of measures depending on concrete textual units.

The corpus used for evaluation was created within Hermes project[1], to evaluate automatic summarizers for Spanish, by comparison to human summarizers. It consists of 120[2] news agency stories of various topics, ranging from 2 to 28 sentences and from 28 to 734 words in length, with an average length of 275 words per story.

To avoid the variability of human generated abstracts, human summarizers built an extract-based golden standard. Paragraphs were chosen as the basic textual unit because they are self-contained meaning units. In most of the cases, paragraphs contained a single

[1]Information about this project available in http://terral.ieec.uned.es/hermes/

[2]For the experiments reported here, one-paragraph news were dropped, resulting in a final set of 111 news stories.

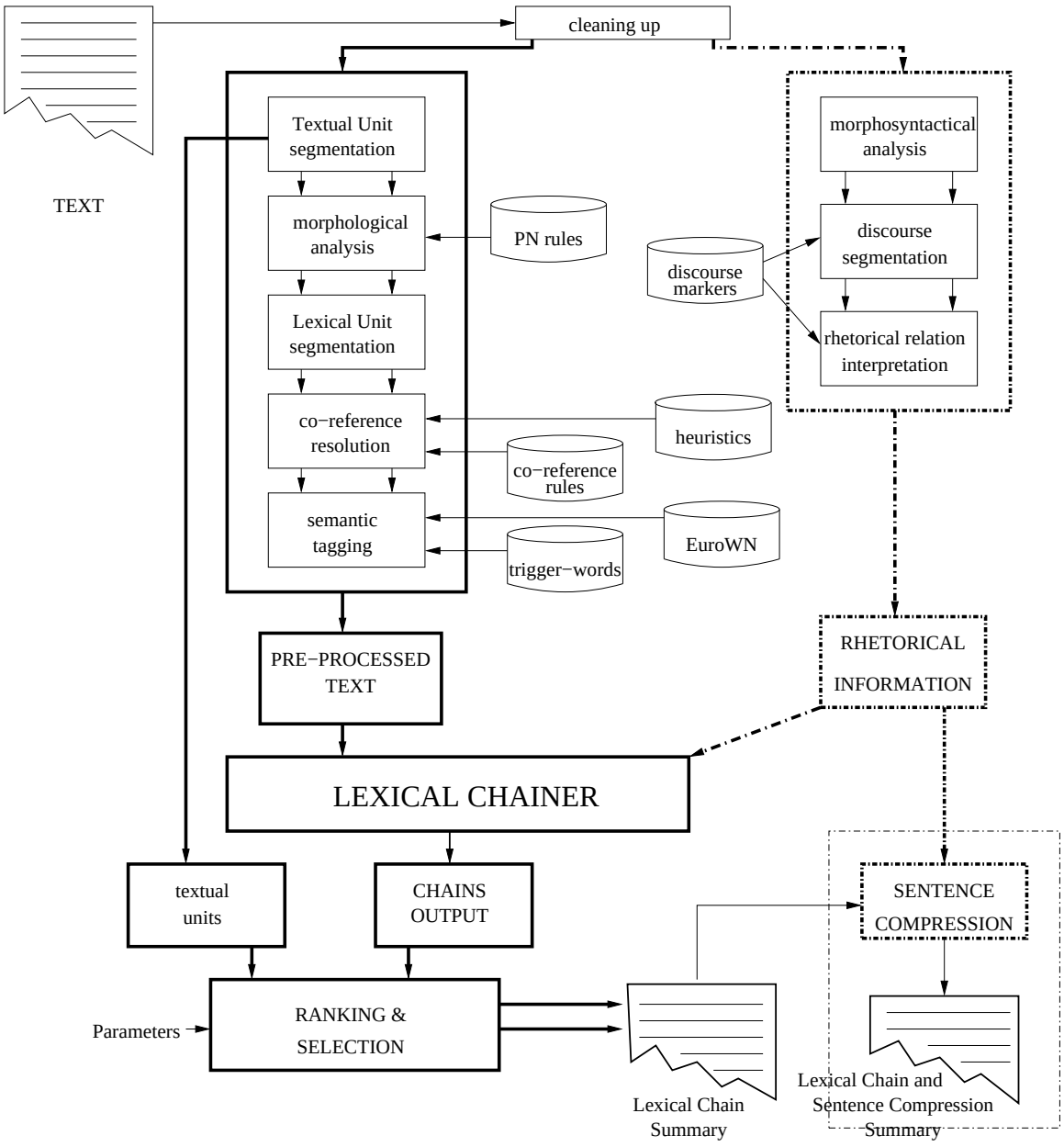


Figure 2.1: Integration of discursive information: lexical chains (left) and discourse structural (right)

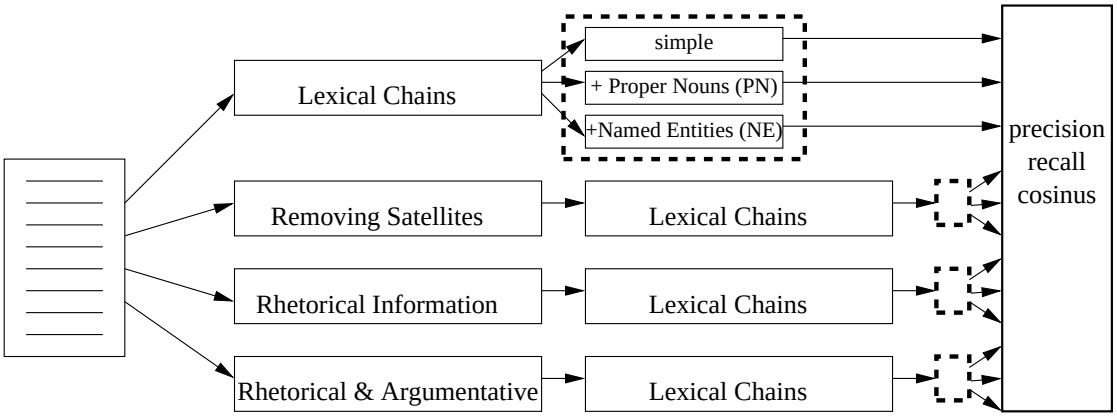


Figure 2.2: Experiments to assess the impact of discourse structure on lexical chain members

sentence. Every paragraph in a story was ranked from 0 to 2, according to its relevance. 31 human judges summarized the corpus, so that at least 5 different evaluations were obtained for each story.

Golden standards were obtained coming as close as possible to the 10% of the length of the original text (19% compression average).

The two main shortcomings of this corpus are its small size and the fact that it belongs to the journalistic genre. However, we know of no other corpus for summary evaluation in Spanish.

The performance of the Lexical Chain System with no discourse structural information was taken as the base to improve. Fuentes and Rodríguez (2002) report on a number of experiments to evaluate the effect of different parameters on the results of lexical chains. To keep comparability with the golden standard, and to adequately calculate precision and recall measures, paragraph-sized TUs were extracted at 10% compression rate.

Some parameters were left unaltered for the whole of the experiment set: only *strong* or *extra-strong* chains were built, no information from defined noun phrases or trigger words could be used and only short co-reference chains were built. Results are presented in Table 2.1.

The first column in the table shows the main parameters governing each trial: simple lexical chains, lexical chains successively augmented with proper noun and co-Reference chains, and finally giving special weighting to the 1st TU because of global document structure appliable to the journalistic genre.

Two heuristics schemata were experimented: *heuristic 1* ranks as most relevant the first TU crossed by a strong chain, while *heuristic 2* ranks highest the TU crossed by the maximum of strong chains. An evaluation of SweSum (SweSum 2002), a summarization system available for Spanish, is also provided as a comparison ground. Trials with SweSum

	Precision	Recall	Cosine
Lead	.95	.85	.90
SweSum	.90	.81	.87
HEURISTIC 1			
Lexical Chains	.82	.81	.85
Lexical + PN Chains	.85	.85	.88
Lexical + PN + coRef Chains	.83	.83	.87
Lexical Chains + PN + coRef Chains + 1st TU	.88	.88	.90
HEURISTIC 2			
Lexical Chains	.71	.72	.79
Lexical + PN Chains	.73	.74	.81
Lexical + PN + coRef Chains	.70	.71	.78
Lexical + PN + coRef Chains + 1st TU	.82	.82	.86

Table 2.1: Performance of the lexical chain Summarizer

were carried out with the default parameters of the system. In addition, the first paragraph of every text, the so-called lead summary, was taken as a dummy baseline.

As can be seen in Table 2.1, the lead achieves the best results, with almost the best possible score. This is due to the pyramidal organisation of the journalistic genre, that causes most relevant information to be placed at the beginning of the text. Consequently, any heuristic assigning more relevance to the beginning of the text will achieve better results in this kind of genre. This is the case for the default parameters of SweSum and *heuristic 1*.

However, it must be noted that lexical chain summarizer produces results with high cosine and low precision, while SweSum yields high precision and low cosine. This means that, while the textual units extracted by the summarizer are not identical to the ones in the golden standard, their content is not dissimilar. This seems to indicate that the summarizer successfully captures content-based relevance, which is genre-independent. Consequently, the lexical chain summarizer should be able to capture relevance when applied to non-journalistic texts. This seems to be supported by the fact that *heuristic 2* improves cosine over precision four points higher than *heuristic 1*, which seems more genre-dependent.

Unexpectedly, co-reference chains cause a decrease in the performance of the system. This may be due to their limited length, and also to the fact that both full forms and pronouns are given the same score, which does not capture the difference in relevance signalled by the difference in form.

2.4.2.1.5 Results of the Integration of Heterogenous Discursive Informations

Structural discursive information was integrated with only those parameters of the lexical chain summarizer that exploited general discursive information. *Heuristic 1* was not considered because it is too genre-dependent. No co-reference information was taken into account, since it does not seem to yield any improvement.

	Precision	**Recall**	**Cosine**
Sentence Compression + Lexical Chains			
Sentence Compression + Lexical + PN Chains	.74	.75	.70
Sentence Compression + Lexical + PN + Chains + 1st TU	.86	.85	.76
Rhetorical Information + Lexical Chains			
Rhetorical Information + Lexical + PN Chains	.74	.76	.82
Rhetorical Information + Lexical + PN Chains + 1st TU	.83	.84	.86
Rhetorical Information + Argumentative + Lexical Chains			
Rhetorical + Argumentative + Lexical + PN Chains	.79	.80	.84
Rhetorical + Argumentative + Lexical + PU Chains + 1st TU	.84	.85	.87

Table 2.2: Results of the integration of lexical chains and discourse structural information

The results of integrating lexical chains with discourse structural information can be seen in Table 2.2. Following the design sketched in Figure 2.4.2.1.4, the performance of the lexical chains summarizer was first evaluated on a text where satellites had been removed. As stated by Brunn *et al.* (2001) and Alonso and Fuentes (2002), removing satellites slightly improves the relevance assessment of the lexical chainer (by one point).

Secondly, discourse coherence information was incorporated. Rhetorical and argumentative informations were distinguished, since the first identifies mainly unimportant parts of text and the second identifies both important and unimportant. Identifying satellites instead of removing them yields only a slight improvement on recall (from .75 to .76), but significantly improves cosine (from .70 to .82).

When argumentative information is provided, an improvement of .5 in performance is observed in all three metrics in comparison to removing satellites. As can be expected, ranking the first TU higher results in better measures, because of the nature of the genre. When this parameter is set, removing satellites outperforms the results obtained by taking into account discourse structural information in precision. However, this can also be due to the fact that when the text is compressed, TUs are shorter, and a higher number of them can be extracted within the fixed compression rate. It must be noted, though, that recall does not drop for these summaries.

Lastly, intra-sentential and sentential satellites of the best summary obtained by lexical chains were removed, increasing compression of the resulting summaries from an average 18.84% for lexical chain summaries to a 14.43% for summaries which were sentence-compressed. Moreover, since sentences were shortened, readability was increased, which can be considered as a further factor of compression. However, these summaries have not

been evaluated with the MEADeval package because no golden standard was available for textual units smaller than paragraphs. Precision and recall measures could not be calculated for summaries that removed satellites, because they could not be compared with the golden standard, consisting only full sentences.

2.4.2.1.6 Discussion The presented evaluation successfully shows the improvements of integrating cohesion and coherence, but it has two weak points. First, the small size of the corpus and the fact that it represents a single genre, which does not allow for safe generalisations. Second, the fact that evaluation metrics fall short in assessing the improvements yielded by the combination of these two discursive informations, since they cannot account for quantitative improvements at granularity levels different from the unit used in the golden standard, and therefore a full evaluation of summaries involving sentence compression is precluded. Moreover, qualitative improvements on general text coherence cannot be captured, nor their impact on summary readability.

As stated by Goldstein *et al.* (1999), "*one of the unresolved problems in summarization evaluation is how to penalize extraneous non-useful information contained in a summary*". We have tried to address this problem by identifying text segments which carry non-useful information, but the presented metrics do not capture this improvement.

We have shown that the collaborative integration of heterogeneous discursive information yields an improvement on the reperesentation of source text, as can be seen by improvements in resulting summaries. Although this enriched representation does not outperform a dummy baseline consisting of taking the first paragraph of the text, we have argued that the resulting representation of text is genre-independent and succeeds in capturing content relevance, as shown by cosine measures.

Since the properties exploited by the presented system are text-bound and follow general principles of text organization, they can be considered to have language-wide validity. This means that the system is domain-independent, though it can be easily tuned to different genres.

Moreover, the system presents portability to a variety of languages, as long as it has the knowledge sources required, basically, shallow tools for morpho-syntactical analysis, a version of WordNet for building and ranking lexical chains, and a lexicon of discourse markers for obtaining a certain discourse structure.

2.4.2.2 Combining heterogeneous information for e-mail summarization

In Alonso *et al.* (2003a), Alonso *et al.* (2003b) and Alonso *et al.* (2004b) we presented Carpanta, an e-mail summarization system that applies a knowledge intensive approach to obtain highly coherent summaries. An on-line demonstration can be found at `http://www.lsi.upc.es/~`
Robustness and portability are guaranteed by the use of general-purpose NLP, but it also exploits language- and domain-dependent knowledge. The system is evaluated against a corpus of human-judged summaries, reaching satisfactory levels of performance.

Carpanta is the e-mail summarization system within project Petra, funded by the Spanish Government (CICyT TIC-2000-0335). The global goal of the project is to develop

an advanced and flexible system for unified message management, which enhances the mobility, usability and confidentiality levels of current systems, while keeping compatibility with main nowadays computer–phone integration platforms. PETRA is related to the European project MAJORDOME - Unified Messaging System (E!-2340), whose aim is to introduce a unified messaging system that allows users to access e-mail, voice mail, and faxes from a common "in-box", and also to make all types of messages accessible directly and/or remotely via Internet or mobile.

The project includes three work lines:

1. **Integration** of phone, internet and fax services, including an e-mail access through HTML client (webmail), which will increase communication mobility and multimedia facilities. This line includes subgoals aiming to the development of a flexible modular architecture, which may present the same information in different forms through different devices (webmail, phone, fax).

2. Development of advanced **oral interfaces** based on speech recognintion and understanding, speech synthesis, and speaker verification.

3. Intelligent **information management** through the use of Natural Language Procesing (NLP) techniques for text classification and summarization, as well as for information retrieval. This task includes the subgoals of advanced Named Entity recognition and correference resolution, document filtering, categorization and retrieval, and text summarization, being this last issue specially relevant for oral interfaces to electronic mail systems.

The summarization module within PETRA is CARPANTA. It is currently working for Spanish, but portability to other languages is guaranteed by its modular architecture, with a language-independent core and separated modules exploiting language-dependent knowledge.

2.4.2.2.1 Problems of e-mail summarization Besides the common problems with the general task of summarization, e-mail summarization has its own idiosyncratic problems:

- noisy input (headers, tags,...)
- linguistic well-formedness is far from guaranteed
- mixed properties of oral and written language
- multi-topic messages

Many scholars have studied relevant aspects of the e-mail register. They have mainly focused on the similarities and differences between oral language and texts (Yates and Orlikowski 1993; Ferrara *et al.* 1990) as well as in brand new intentionally-expressive devices, such as previous-message cohesion (Herring 1999), visual devices (Fais and Ogura 2001), simplified registers

(Murray 2000) or internet-users vocabulary. Nevertheless, they disregard a factor that is important in the e-mail register: as the user often writes fast and not much reflectively, texts contain many non-intentional language mistakes.

Nevertheless, in our opinion, those studies do not systematize fully the problem as they disregard a factor that is important in the e-mail register: as the user often writes fast and not much reflectively, texts contain many non-intentional language mistakes. Other factors which are not regarded either, probably because most studies have been carried out in monolingual English-speaking environments, are other kinds of intentional non-standard performance, such as systematic lack of accents, or language shifts or interference.

In a recent study, (Climent *et al.* 2003) argue that, for their universe of study, more than 10% of the text in emails are made of either non-intentional errors, intentional deviations of the written standards, or specific terminology. For Spanish, 3.1% of the words contain either performance or competence errors, another 3.3% are either language-shifts or new forms of textual expressivity (such as ortographical innovations or, specially, systematic non-accentuation), and another 4.2% consist of specific terminology -thus words usually missing from many system's lexicons.

This supposes two important drawbacks for our purposes: (i) the task of the analyzing tools is severely endangered as they should deal with an extremely noisy input; and (ii) the eventual speech-synthesized output will also be poor or even in part hardly understandable. In any case, such a bulk of asystematic differences from standard texts implies a barrier for high-quality, general-purpose NLP tools. As a consequence, very little work has been done on quality e-mail summarization. (Tzoukermann *et al.* 2001) aim to capture the gist of e-mail messages by extracting salient noun phrases, using a combination of machine learning and shallow linguistic analysis.

2.4.2.2.2 Approach Considering e-mail summarization problems and the environment within PETRA project, summaries produced by CARPANTA have the following properties:

oral output by telephone,

indicative summaries just give a hint of the content, to meet the severe restrictions of length imposed by the oral format,

coherent because the summary cannot be revised as easily as written ones, (thus excluding list-of-words approach),

extractive due to limitations for general-purpose NLP tools,

knowledge-intensive combining analysis at different linguistic levels, IR techniques and IE strategies specific for e-mail, in order to build a robust system that is also capable of producing deep analyses.

2.4.2.2.3 Architecture of the System As can be seen in Figure 2.3, CARPANTA is highly modular, which guarantees portability to other languages.

E-mail specific knowledge has different status within the system, so that language-dependent modules can be updated and switched to address concrete necessities (different

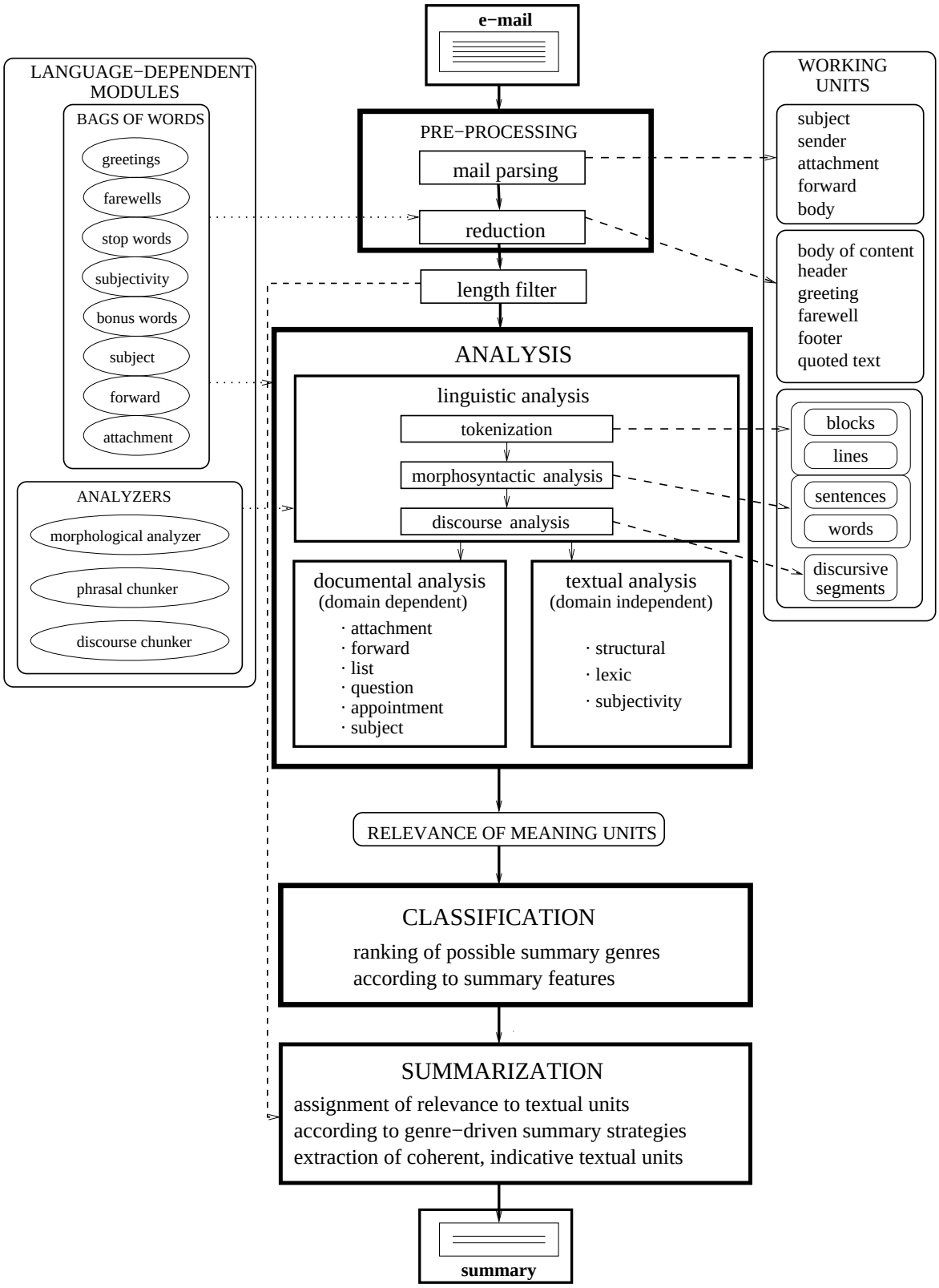


Figure 2.3: Architecture of CARPANTA.

languages, restricted domains), while language-independent strategies form part of the core processing stream. In addition to general-purpose NLP tools, the following e-mail specific resources were developed:

- a classification where each kind of e-mail is associated to its most adequate summary and summarization strategy (language-independent) (seen in Table 2.3)
- bags of words and expressions that signal different kinds of e-mail specific contents (language-dependent):
  - greetings, farewells,
  - reply, forward, attachment
  - bags of words signalling different kinds of relevance: personal involvement of the writer in the message, information exchange; also lack of relevance.
- strategies to deal with these anchors and their associated content (language-independent)

The process for e-mails to be summarized is described in what follows, and exemplified in Figure 2.5.

Parse e-mail format. Messages undergo a pre-processing to identify headers, greetings, visit cards, quoted text, and the body of text, which is further analyzed.

Linguistic analysis. First, the body of text is analyzed morphosyntactically (Atserias *et al.* 1998a) and chunks are identified (Atserias *et al.* 1998b). Then, discourse chunks, signalled by punctuation and discourse markers, are found (what we call *segments*). Finally, the salience of non-empty words is calculated according to the frequency of occurrence of their lemma.

Textual analysis. Three different kinds of textual relevance have been distinguished: lexic, structural and subjective. For each of these three aspects of e-mails, a global reliability score is obtained, taking into account how well each kind of information distinguishes relevant and non-relevant pieces of the e-mail. Then, relevance is also calculated with respect to meaning units, basically, discourse segments. Lexic relevance of a segment is directly proportional to the amount of frequent words in the segment and inversely proportional to the length of the segment. Structural relevance is assigned as a result of the interpretation of discursive relations between segments and between a segment and the whole text, by means of the information associated to discourse markers. Finally, subjective relevance is found when the segment contains any of a list of lexical expressions signalling subjectivity.

Documental analysis. Key words and expressions signalling information specific of e-mail (e.g., appointment, list, etc.) are detected by simple IE techniques, basically, pattern-matching.

As a result of linguistic, textual and documental analysis, a set of meaning units is produced at different linguistic levels: words, chunks, segments and sentences, but also lines and paragraphs. Each unit is assigned a complex relevance score, one for each kind of information that is taken into account. Values for lexical, structural and subjective relevance are continious, ranging from 0 to 1. Each unit is also assinged a binary relevance

score for each kind of e-mail specific information, 1 if there is any clue signalling that kind of information in the unit, 0 if there is none.

Classification The most adequate summarization strategy is determined by taking into account the characterizing features of each e-mail, as provided by the analysis module. The general schema followed by classification rules can be seen in Figure 2.4, Table 2.3 shows the relation between e-mail features and summarization strategies.

Summarization Then, the chosen summary is produced. Different kinds of summaries are described in Table 2.3.

summarization approach	**summary**	**textual features**	**documental features**
full mail	whole e-mail text	short (<30 words)	
pyramidal	first compressed paragraph	–	–
lead	first compressed sentence	–	–
subject	subject of e-mail	lexical relevance	subject is relevant
appointment	seg. stating appointment	–	appointment
attachment	seg. describing attachment	–	attachment
forward	seg. describing forward	–	forward
question	seg. with question	–	question mark
list	seg. preceeding the list	–	list
lexic	seg. with most relevant lexic	lexical relevance	–
structural	most structurally salient seg.	structural relevance	–
subjective	seg. with subjectivity	subjective relevance	–
textual	most relevant seg. summing textual evidence	–	–
textual + documental	most relevant seg. summing text and documental evidence	–	–

Table 2.3: Classification of summaries, characterizing features and summarization strategies.

2.4.2.2.4 Evaluation To tune and evaluate the performance of the system, the automatic summaries produced were compared with summaries produced by potential users of the system. 200 e-mails were summarized by 20 judges, so that each e-mail was summarized by at least 2 judges. To produce a summary, judges selected those words in the e-mail that they would consider a satisfactory summary of the message if they received it by telephone. The corpus of original and summarized e-mails is available at `http://www.lsi.upc.es/~bcasas/carpanta/`. 20% of the judged e-mail was left for evaluation, the rest was used for tuning the system.

The *kappa* measure (Landis and Koch 1977) was used to calculate pairwise agreement between judges. Kappa is a better measurement of agreement than raw percentage agreement because it factors out the level of agreement which would be reached by random.

```
if strong genre evidence
   if strong linguistic evidence
       textual + documental
   else
      if evidence for a single genre
          specific strategy
          (list, question, attachment)
      else combination of genres
else
    if strong textual evidence
        textual
    else lead
```

Figure 2.4: General schema followed by classification rules.

When there is no agreement other than what would be expected by chance, $k = 0$, when agreement is perfect, $k = 1$.

The obtained kappa values for agreement between judges ranged from 0.36 to 1, with a mean of 0.75 and a standard deviation of 0.17. Following (Carletta 1996), we can consider that kappa values above 0.7 indicate good stability and reproducibility of the results, so it can be said that it is possible to discriminate a good e-mail summary from a bad one, and that it is even possible to determine the best summary for a given e-mail.

The goodness of automatic summaries was calculated as the agreement with the corresponding human summaries. As a global measure of the system's performance, we calculated the effect of considering the system as a human judge more, with respect to average kappa agreement. Taking the 20% of the corpus left apart for summarization, we obtained that the average kappa agreement between human judges was 0.74, and it decreased to 0.54 when the system was introduced as a judge more. This indicates that the system does not as well as human judges, but still, a kappa value bigger than 0.4 indicates moderate agreement.

2.4.2.2.5 Results and Discussion Figure 2.6 shows the results of comparing automatic summaries against human-made summaries of the 40 e-mails reserved for evaluation. For each e-mail, automatic summaries were obtained using all of the summarization strategies applicable, for example: *lexic*, *structural*, *appointment*, *attachment*, etc. Then, kappa agreement was calculated between each automatic summary and every human summary provided for that e-mail.

To tune and evaluate the performance of the system, the automatic summaries produced were compared with summaries produced for 200 e-mails by 20 potential users of the system, with a minimum of 2 different human summaries for each e-mail. Agreement between judges ranged from $\kappa = -.37$ to $\kappa = 1$, with a mean of $\kappa = .47$, which indicates

Original Message:
Hola Laura,
No te preocupes por el seminario de la semana que viene. Ya hemos encontrado un articulo y se lo hemos asignado al Oriol. Vamos a hacer una introducción a los HMM, que hay mucha gente que no sabe como funcionan y estan por todos lados. El articulo es muy introductorio, me parece, y si quieres te guardo una copia. Por lo del network de discurso: me parece fantastico. Si pasas por aquí podemos hablar un rato. Estoy en el despacho 300B y mi extensión es 2234. No se si pasas por aqui a veces. De todos modos hay más gente a que le va interessar el tema de discurso. En principio estoy aquí todas las mañanas y casi todas las tardes (menos los martes), pero por las tardes siempre hay alguna historia. Si quieres nos podemos encontrar un dia y hablar un rato. Yo no he hecho mucha cosa aún. Estoy empezando a buscar cosas y leer un poco, pero aún no tengo ninguna idea en concreto. Quizás me puedes recomendar algo que sea introductorio y que se lea rapido para entrar en la tematica.
Hasta luego,
Stefan

Relevant segments, with their relevance features:

- El artículo es muy introductorio, me parece, y si quieres te guardo una copia. (*The paper is very introductory, it seems to me, and if you want I can save you a copy.*)
  - structural relevance 0.5
  - subjectivity relevance 0.5 (*parecer*, *"seem"*; *introductorio*, *"introductory"*)
- Si pasas por aquí podemos hablar un rato. (*If you pass by we can talk for a while.*)
  - structural relevance 0.5
  - lexical relevance 0.8 (*pasar*, *"pass"*; *aquí*, *"here"*; *hablar*, *"talk"*;*rato*, *"while"*)
- Si quieres nos podemos encontrar un día y hablar un rato. (*If you want we can meet some day and talk for a while.*)
  - structural relevance 0.75
  - lexical relevance 0.5 (*hablar*, *"talk"*;*rato*, *"while"*)
  - subjectivity relevance 0.5 (*quieres*, *"want"*, *podemos*, *"can"*)
  - evidence of appointment (*encontrar*, *"meet"*)

Chosen Summary (11 words)

Si quieres nos podemos encontrar un día y hablar un rato.

Figure 2.5: Example of the summarization process by CARPANTA.

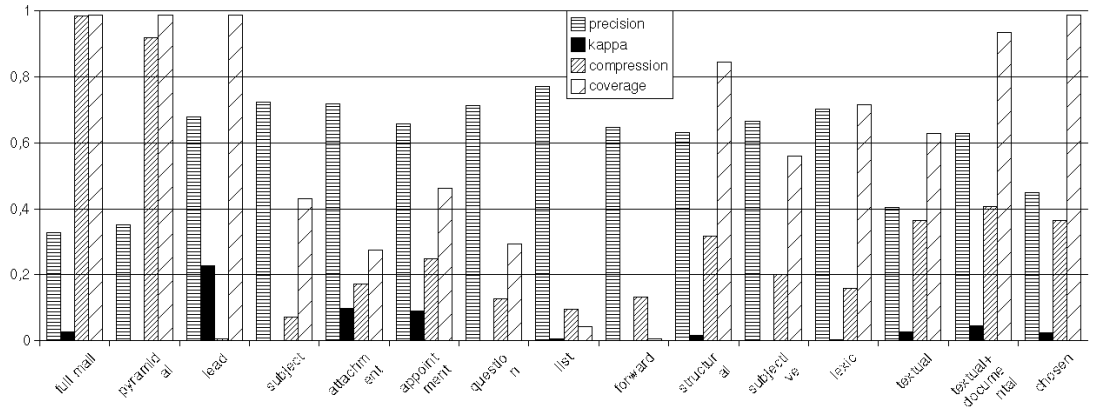


Figure 2.6: Main features of the performance of different summarization strategies: average length with respect to the original e-mail text (compression), coverage over the test collection, and kappa agreement and precision by comparison to human summaries.

that agreement is far beyond chance, but also that the task of e-mail summarization is somewhat fuzzy for users.

The goodness of automatic summaries was calculated by comparison with the corresponding human summaries, results can be seen in Figure 2.6. For each e-mail, automatic summaries were obtained using all of the summarization strategies applicable, based on linguistic information (*lexical*, *structural*, etc.), on e-mail specific information (*appointment*, *attachment*, etc.) in both (*textual and documental*) or applying baseline strategies, like having the first line or paragraph as the summary.

Human and automatic summaries were compared by κ agreement and by precision at discourse unit level. Agreement between human and automatic summaries was very low in terms of κ (average $\kappa = .02$), but evaluation metrics more usual for summarization, like precision with respect to human summaries, reached 60% average, which is the state of the art for automatic text summarization.

Results show that simple methods, like taking the first line of the e-mail (*lead*) offer very good results, but, in general, summaries exploiting e-mail specific knowledge (*list, appointment*) can improve on this baseline. However, these kinds of e-mail present very low coverage. The strategy combining general linguistic and e-mail specific knowledge (*textual and documental*) yields a good balance between coverage and precision.

Finally, results concerning the chosen summary show that there is still room for improvement within the classification module, since most of the alternative summaries present higher precision rates than the chosen one.

2.4.2.2.6 Conclusions and Future Work We have presented CARPANTA, an e-mail summarization system that applies a knowledge-intensive approach to obtain highly coherent summaries, targeted to guarantee understandability in delivery by phone. The performance of the system has been evaluated with a corpus of human-made e-mail sum-

maries, reaching a level of agreement with users close to agreement between human judges. However, results indicate that the classification module has to be improved, which will be done by manually incrementing the rules and by applying machine learning techniques.

Given the highly modular architecture of CARPANTA, adaptation to other languages has a very low cost of develpment, provided the required NLP tools are available. Indeed, enhancements for Catalan and English are under development.

Future work in this system should include modules that enable for automatic normalization and correction of input texts. (Climent *et al.* 2003) suggest that there's special need for modules of: (a) punctuation recovery, (b) accent recovery, (c) spelling-mistake correction, and (d) terminological tuning according to users' profiles.

In this section we have argued that a certain representation of the discursive structure of texts can contribute to the improvement of the performance of AS systems. Moreover, we have argued that such representation could be obtained by exploiting shallow textual cues. We have supported this position by a review of previous work and also by analyzing in detail the performance two systems that exploit discursive information, and showing how this kind of information effectively contributes to the quality of the resulting summaries.

2.5 Discourse for text summarization via shallow NLP techniques

In this section we describe our approach to obtain a representation of the discursive organization of texts that is useful for text summarization and can be obtained by shallow NLP techniques.

The aim of our analysis is to identify content units at discourse level and the relations between them that allow to determine their relevance within the text and the requirements for the final summary to be coherent, in terms of the configuration of discourse units that are needed.

A very important aspect of our approach is that we want to obtain summaries not only for English, but also for other languages. We have focussed in Catalan and Spanish, but the approach we present is arguably valid for languages with limited NLP resources in general. Then, discursive analysis is based on the capabilities of the available NLP tools for Catalan and Spanish. As has been argued in the previous section, this supposes a limitation, but also an interesting perspective about the analysis of texts. Indeed, the obtained representations are data-driven and, as a natural consequence, the theory to which these evidence may be related has a strong empirical basis.

2.5.1 Delimiting shallow NLP

Within AI and NLP, discourse has been mainly approached from knowledge-rich perspectives, resulting in theoretical proposals (Hobbs 1985; Polanyi 1988; Polanyi 1996), very restricted domains (Schank and Abelson 1977) or very restricted aspects of discourse

(Hirschberg and Litman 1993). However, following the general interest for robust NLP systems, the interest for robust discourse parsing has also changed perspectives in discourse processing. More recent systems have shifted the emphasis from completeness to coverage, which provide shallow analyses but are capable of tackling a wider spectrum of cases, as has been explained in Section 2.4.

From what has been presented in the previous section, it seems clear that the representation of the discursive aspect of texts can significantly improve the quality of resulting summaries, even if this representation is obtained by shallow NLP techniques. We have also argued that a shallow approach guarantees the robustness of the approach, and that it can be very informative of the organization of texts provided it is related with a theory of textual organization in a principled way.

A shallow approach linked to a theory guarantees that the analysis procedure will fail only exceptionally, and that some interpretation of the text to be analyzed will be provided in most cases, even if this analysis is not very informative. This is due to the fact that the most basic principles of discursive structuring can be found in every textual document. This makes it possible to hypothesize a default representation of discourse structure to explain any document. This default can be improved or enriched if clues are found that allow to drive more informative conclusions on the discursive organization of a given document, and these clues are related to a theory of textual organization in a principled way. The analysis of these clues can benefit from knowledge-rich, deep reasoning techniques, but it is also possible to carry it out with shallow techniques, as we have shown with the applications in Sections 2.4.2.1and 2.4.2.2.

To our knowledge, the current capabilities of NLP tools available for Catalan and Spanish are: tokenization, stemming, morphological analysis and disambiguation and chunking of simple phrases. All of them can be carried out with the FreeLing NLP toolset (FreeLing), an open source language analysis tool suite that provides this basic linguistic analysis for Catalan, Spanish and English. Deeper analyses, like full syntactic parsing, word sense labelling or propositional representation cannot be obtained with a reliable degree of certainty for Catalan or Spanish, so we have taken the output of FreeLing as the basis for discursive analysis.

There is a wealth of linguistic information that has been recognized as influencing the configuration of discourse, like information packaging, argumental schemata, syntactical functions or semantic content, including lexical semantics but also the resolution of referential expressions, presuppositions or implicatures. However, given the current limitations of the NLP tools and resources for Catalan and Spanish, we cannot exploit information that requires more than a partial syntactical analysis. Nevertheless, we resort to this kind of information to fully characterize the discursive phenomena from a theoretical point of view in Chapters 3 and 4.

Since discourse units and their relations are identified and characterized by simple techniques, the information provided by **shallow textual cues** is crucial. With the NLP capabilities described above, we can exploit: punctuation, discourse particles, some syntactical structures, specially conjunctive ones, patterns of lexic and referential expressions, and, in some languages, also word order. But the most useful of all shallow clues available

in a text are **discourse markers**, because they are highly informative of semantically rich relations between discourse units, because they can be satisfactorily associated with a finite set of processing instructions that explicitly signal their contribution to the structure of discourse. This limited amount of information constrains the kind of representation that can be obtained, as is detailed in the specification of our targeted representation, described in Chapter 3.

Although we aim to connect shallow evidence with theoretical insights on the organization of texts, we are not commited to a specific theory. Some of the discourse parsers discussed in Section 2.4 rely on well established theories of discourse or some adaptation of them. For example, Corston-Oliver (1998), Marcu (2000) and Soricut and Marcu (2003) implement RST, and Schilder (2002) makes an adaptation of SDRT. However, the DL-TAG (Forbes, Miltsakaki, Prasad, Sarkar, Joshi and Webber 2003) does not rely upon a well-established theory, but builds upon empirical findings systematized in different studies. We are also ecclectic as to our underlying theoretical model, which will be explained in the following two chapters.

2.5.2 General approach to obtain a representation of discourse

In our analysis, we priorize precision over coverage and informativity, so that it can be guaranteed that every analysis is true with a reliable degree of certainty. As for coverage, we provide a default representation for all cases, so that none is left uncovered. But this representation is only as informative as the available evidence allows. Therefore, informativity is the varying parameter in our analysis scheme: the more evidence, the more informative the analysis will be, the less evidence, the less informative. As will be explained in the following chapter, this causes a conflict with the requirement that the representation of discourse is homogeneous, as proposed by some theoretical approaches.

The basic procedure of analysis at discursive level follows Marcu (2000), as follows:

1. identify the textual evidence of structure that applies for the current level of analysis, and determine its function in case they are ambiguous,
2. determine the relative confidence of the textual evidence found, so that decisions can be made in case different evidence leads to conflicting analyses,
3. determine discourse units at that level, probably building upon the result of the analysis at the previous level,
4. determine relations between discourse units.

In the next chapter we present a specification of the representation of discourse that we are aiming for. We present the starting assumptions of our work, and describe the basic concepts of the representation. As in Polanyi (1996), we distinguish three basic elements to be taken into account in a model of discourse (in our case, low level discourse): the structure of discourse or theory of the organization of units, the discourse units that compose it, and a special kind of discourse units that provide information about the relations of units within the structure, the so-called *discourse markers*.

2.6 Discussion

In this chapter we have analyzed the field of automatic text summarization (AS), highlighting open issues like evaluation and the properties that characterize an insightful, scalable approach to summarization.

Concerning evaluation of summaries, we have discussed that well-established methods do not provide a good assessment of summary quality, because they are based on the comparison with a single summary considered as a *gold standard*, even if it has been widely acknowledged that a unique gold standard cannot capture the variability that is intrinsic to the summarization task. We describe some recent methods that try to address this problem, and we argue that they could be enhanced by an automated analysis of the source text as the one we propose in this section.

Then, we make a brief overview of previous work in the area of AS. We find that satisfactory approaches to AS exploit general properties of documents to guarantee robustness of the strategy, while leaving room to gain deeper insight into the configuration of texts if richer information is available. We have found that shallow approaches are interesting because they seem to provide results of a quality comparable to that of approaches exploiting deep knowledge techniques, while having wider coberture, and they are more flexible to treat unseen cases.

We have argued that a representation of the organization of texts at discourse level is useful for text summarization, and we have provided various examples of it, including our own work in the integration of discursive analysis in two AS systems. We have determined that such a representation should identify content units at discourse level, their relative relevance in the text and the coherence relations they establish with other units. Then, we have sketched our approach to obtaining a representation of discourse that relies on shallow NLP techniques, more concretely, those that provide a reliable analysis for Catalan and Spanish.

This approach is discussed in the following two chapters: Chapter 3 describes the formal structure by which we represent discourse, the nature of discourse units and how they can be characterized via shallow NLP techniques. Chapter 4 deals with the meaning of discourse relations, delimiting what can be obtained by shallow NLP, proposing a method to induce meanings from textual evidence at surface level and establishing an inventory of discourse meanings that is useful to identify and characterize relevance and coherence relations between content units in a text.

Chapter 3

Specifying a representation of discourse

From what has been exposed in the previous chapter it seems clear that a certain representation of discourse may improve automatic text summarization. This representation should identify functional units at discourse level and also their role in the text, that is, how they are related to other units. Since we want to exploit this information for text summarization, we will focus on coherence and relevance relations, leaving any other kind of relations aside.

Representations of discourse have been typically obtained by resorting to rich analyses of language, involving deep syntactic and semantic analysis and even some kind of reasoning. But all this is well beyond the scope of our current NLP capabilities. Given our framework, we obtain this representation of discourse by exploiting shallow cues via shallow techniques. As will be discussed in the following chapter, the information about discursive structure provided by this kind of evidence is reliable only at a short scope, therefore, only local representations of discourse can be reliably obtained. This constraint has determined that we represent discourse as a sequence of local structures and that different dimensions of meaning are described independently.

In this chapter we will characterize a computational structure to represent relevance and coherence relations in discourse, focusing on those relations that can be reliably obtained by shallow NLP techniques. The linguistic aspects of discourse relations are discussed in the following chapter. There, we present the shallow textual clues that we exploit to identify discourse relations in text, we delimit the kind of relations that we can reliably obtain from them and we describe the scope and meaning of these relations.

The structure of the chapter is as follows. First of all, we elicit our assumptions on the organization of discourse. Then, we delimit the structure by which we represent discourse in Section 3.2. After that, discourse units are characterized, trying to capture theoretical insights in our computational framework. We describe discourse segments in Section 3.4 and discourse markers in Section 3.3.

3.1 Assumptions about the organization of discourse

First of all, it is a fundamental assumption in the area of computational discourse processing that discourse is **coherent**. But what does it mean that discourse is coherent?

Following the cooperation principle (Grice 1969), in coherent discourse the speaker tries to convey relevant information in linguistic form so that the hearer can retrieve the relevant information therefrom. In order to guarantee that this communicative process is successful, the message has to conform to certain conventions on the encoding of information that are shared by both participants. It is by exploiting these conventions, reflected in linguistic form, that we can obtain a representation of the meaning shared between the participants.

Just as the conversational maxima that follow from the cooperation principle, there are some **principles of textuality** that must hold for discourse to perform its communicative function properly. We are not explicitly adopting a concrete set of principles, but we will have this concept, and subsets of it, available as an axiom to explain why coherence holds or should hold at given spots in texts.

Even if we do not go into detail into the principles of textuality, it must be noted that there are various well-established accounts of them in the literature. For NLP, a reference work on the principles of textuality is Halliday and Hasan (1976), who define *cohesion* as *the set of possibilities that exist in the language for making text hang together* (Halliday and Hasan 1976), and distinguish four of such possibilities: *reference*, *substitution*, *ellipsis*, *conjunction* and *lexical*. Any or all of these mechanisms, or all of them may be exploited to configure the discursive structure of text.

A usual assumption in discourse processing is that discourse is **segmented**, so that distinct units can be found in a discourse. It is a property of coherent discourse that **every discourse unit is related** to at least one other discourse unit or construct, like intentions, plans, etc.

Discourse processing can be modelled as an **incremental** task, where higher-level representations are built compositionally upon lower level ones. The form of linguistic utterances is systematically associated to some meaning. When a form has been associated to a meaning, it creates a new form that is available for the next level of processing. Therefore, **distinct levels** can be distinguished in discourse.

Discourse levels tend to be **multidimensional**, that is to say, there are heterogenous organization mechanisms that account for part of the organization of discourse within a single level. These mechanisms usually interact, but an autonomous account of each of them can be pursued, just like the phonological, morphological, syntactic and semantic levels of linguistic structure: with relative autonomy but also relative dependency of each other.

Every discourse unit has a **relative salience** within discourse. Psychological experiments (Kintsch 1988; Sanders and Spooren 2001) have shown that salient discourse units are available for further discourse, while non salient units are not. The relative salience of each unit should be equivalent to some function of its discourse features.

3.2 A structure to represent discourse

3.2.1 Delimiting a representation of discourse

Discourse representation has often been addressed as a knowledge-rich task. As a consequence, most of the approaches to represent discourse are based on a series of very strong underlying claims, which are well beyond the capabilities of shallow NLP.

For example, it has often been claimed that discourse relations hold even if they are not overtly marked by any linguistic mechanism. However, with shallow NLP techniques we cannot identify a discourse relation unless there is a textual clue tht overtly signals it. Therefore, we will establish content-rich relations when explicit textual clues signalling them can be found, and default relations elsewhere.

Default relations are defined precisely as those relations that hold whenever no evidence can be found to support the existence of any other relation. Therefore, they the least marked case of the range of possible analyses for a given case, and so they constitute the basis for an organization of discursive meaning in a range of markedness where defeasible inference can be naturally applied, as has been proposed both in formal discourse analysis and shallow NLP. From the point of view of application, the fact that a default representation is always available guarantees the robustness of a system, because every text can be analyzed, even if the analysis provides only trivial information about the organization of the text.

It is also a common claim that a text should be described by a single structrue encompassing the whole of it. Some approaches even claim that this structure must be homogeneous, that is, that the structure must be constitued by relations of the same kind, and, more importantly in our case, equally informative. For instance, if a text is represented by a hierarchical structure, the same set of relations must be able to explain relations between discouse units at different levels in this structure. In the case of shallow NLP, it is possible to identify content-rich relations between discourse units, but the kind of evidence whereupon we rely to identify these relations does not occur homogenously in all the text and it is only reliable at a short scope. If a text is to be fully covered with a homogeneous structure, content-rich relations cannot be represented, precisely because they are not homogeneous. If the requirement of homogeneity is discarded, a text can be fully covered by semantically heterogeneous relations, in our case, semantically rich relations vs. default relations.

Many structures have been proposed to represent the organization of discourse. A classical view, directly rooted in the field of representation of knowledge, proposes to represent texts as instances of pre-defined frames (Schank and Abelson 1977; McKeown and Radev 1995; Kan 2003; Teufel and Moens 2002a), where each unit is defined by its role in the frame (introduction, initiator, etc.)

Another approach to the organization of discourse that has been exploited by many researchers is that proposed by Grosz and Sidner (1986). They represent discourse as a dynamic structure, more concretely, as a stack where discourse units are popped or pushed depending on the topical progress of the text, so that the topic of the discourse in the

moment m is the topmost element in the stack at moment m.

A number of researchers have proposed that discourse should be described as a tree-like structure where discourse units are represented as nodes and discourse relations are represented as edges linking these nodes (Hobbs 1978; Hobbs 1985; Mann and Thompson 1988; Polanyi 1988; Webber 1988). In this kind of structure, proximity to the root implies some kind of relevance or centrality to the topic of the discourse. But it has been argued that a tree-like structure fails to capture configurations of relations like those presented in 3.1, and that a structure more expressive than a tree would be needed to represent these configurations, like a directed acyclic graph.

Our proposal is that discourse is represented as a kind of highly restricted directed acyclic graph. We say that it is restricted because it is formed by trees, which represent local structures, and which are linked with each other by non-hierarchical relations. Thus, the DAG we need to represent discourse is, properly speaking, a list of trees. Trees represent local structures constitued by hierarchical content-rich relations between discourse units, obtained in the presence of explicit textual clues. Thus, these local structures are delimited by the absence of such clues.

We consider that different dimensions of discursive meaning can be distinguished. These dimensions are represented by different trees covering the same set of terminals. The whole range of meanings in a dimension may hold between segments within a local structure, but structures are liked to each other only by default relations in each dimension. Default relations are the least marked case in the range of meanings in a dimension.

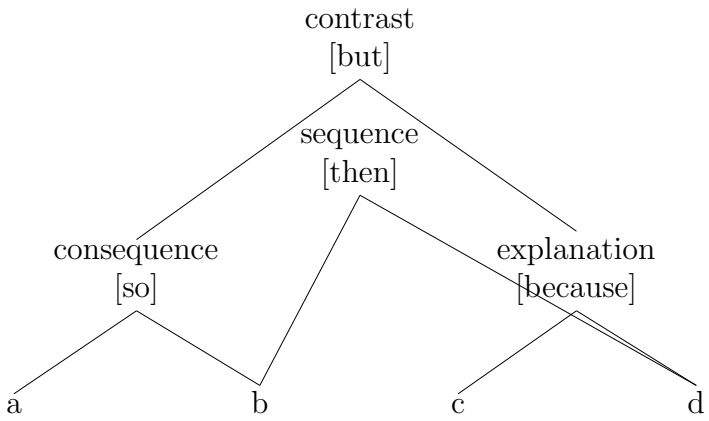
As follows, in our approach the whole text is covered by a single structure, but this structure may be considered non-homogeneous from the point of view of informativity of the relations. Indeed, default relations are qualitatively different from content-rich ones, since they are less informative. Thus, this representation is less informative than the traditional trees covering a whole text and leading to a single root, but it seems to capture a certain aspect of the organization of texts, namely, that texts can be regarded as a sequence of topic-like slots occupied by units that present a content-rich internal organization, as argued by Knott *et al.* (2001). Moreover, being a restricted DAG, its processing cost is rather low.

Representing heterogeneous meaning in different dimensions allows to explain configurations of discourse relations like that presented in Figure 3.1, which cannot be adequately captured by a strict tree-like structure. As is presented in 3.1-(1), the relations between nodes $d \rightarrow b$ and $d \rightarrow c$ can only be captured by crossing branches. To explain the relations established in the text of Figure 3.1 without crossing branches, Webber *et al.* (2003) have proposed treating the relations introduced by anaphoric discourse markers like *then* as qualitatively different relations, resulting in the structure displayed in 3.1-(2). We propose that heterogeneous meanings are represented independently, by different trees, as in 3.1-(3). Besides overcoming the problem of crossing branches, such multidimensional representation seems very adequate to capture the idiosincracies of discursive meaning, as we have already pointed out before and will develop in the following chapter.

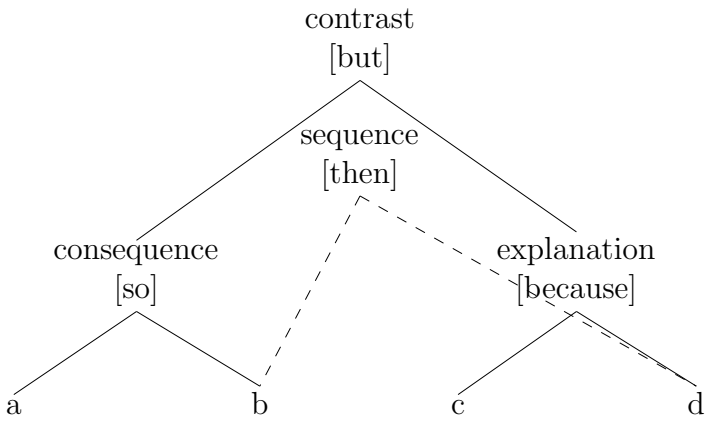
Although in our theoretical model we will always assume that heterogeneous dimensions of meaning can be represented by different trees, implementations for shallow NLP do not

a. John loves Barolo.
b. So he ordered three cases of the '97.
c. But he had to cancel the order
d. because *then* he discovered he was broke.

(1)



(2)



(3)

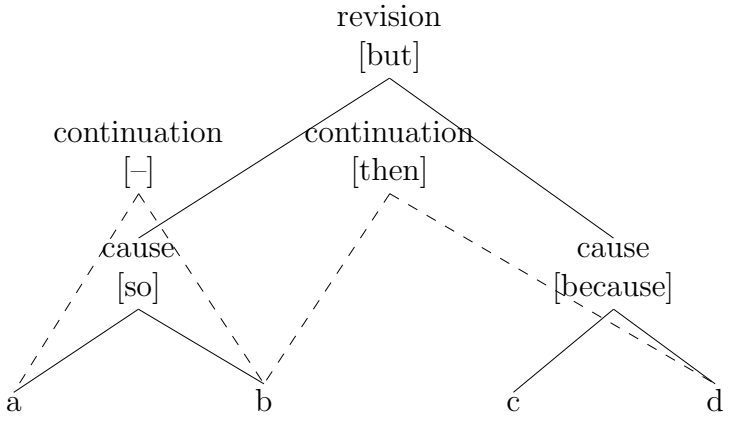


Figure 3.1: Discursive relations that have to be expressed by crossing branches in a strict tree-like representation, and alternative approaches proposed to explain this configuration of discourse relations without crossing branches: as anaphoric relations (2) or by distinguishing different dimensions of meaning (3).

have a need to resort to a multidimensional tree. Indeed, the information provided by shallow evidence does not lead to a configuration of relations that needs to be represented by crossing branches in a strict tree-like structure. As said before, crossing branches are proper of discourse markers with anaphoric properties. These properties cannot be properly treated by shallow NLP techniques because they require knowledge-rich processing and because they typically deal with long-scope relations, which we have left aside. We represent discourse as a list of unidimensional trees, but represent relations between nodes in a tree so that they can be easily translated to multidimensional trees.

3.2.2 Specification of the structure

As we have said, we will represent the structure of discourse as a sequence of trees, without any restrictions as to their dimensionality.

The properties of the proposed structure are (the terms *node* and *discourse unit* are used indistinctly, as are the terms *relation* and *edge*):

uniqueness each discourse unit is related to only one discourse unit in each of the possible dimensions of meaning. The set of discourse units is the same for each dimension of meaning.

full coverage each discourse unit is related to another discourse unit in each dimension of meaning, even if it is related only by the default relation[1]. The only exception is the root of the tree (in our case, the root of the first tree in the list of trees), which is attached to itself, and to which subsequent nodes attach. As follows, the structure of discourse covers the whole text.

labelled edges are labelled, even if they are labelled by a default meaning.

binary relations are always established between pairs of discourse units conveying propositional content (discourse segments). A discourse unit that conveys only procedural content (discourse marker) functions as an edge and label between a pair of discourse units other than itself. We argue that there is no need to have n-ary relations for descriptive adequacy, because the so-called *multinuclear* relations, where the same relation is said to hold for an arbitrary number of units (for example, lists), can be well described as a concatenation of binary relations, where each unit is related to the unit immediately preceding it.

directionality and asymmetry the edge relating a pair of nodes is directed, departing from the discourse unit that is marked for a given relation. For example, if two discourse units are related, the one that is marked for a given relation is the one dominated or containing the textual clue that signals this relation (discourse marker,

[1]For a detailed description of default relations, see Section 4.2.3 that account for the content conveyed in each dimension.

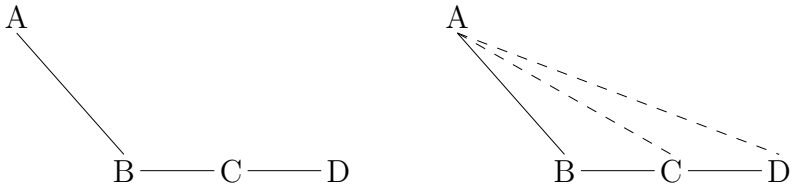


Figure 3.2: Applying regular inference processes associated to the default relation of *continuation*, a strict tree-like representation (left) can be transformed in a DAG (right).

syntactical structure, position in the sequence of text[2], etc.). This allows that a marked discourse unit can only bear a relation with one other unit in a given dimension of meaning, but many units can establish a relation with an unmarked unit, as can be seen in the following example, where the unmarked unit, the main clause "*Father Thadeus Nguyen Van Ly was arrested*" establishes a many-to-one relation with the marked units (b, c, d), which in turn establish only one relation each, with the main clause. Note that the directionality of relations is not necessarily mapped to the topographical configuration of the structure of discourse.

(4) [$_a$ Father Thadeus Nguyen Van Ly [$_b$, a Roman Catholic priest in Viet Nam, $_b$] was arrested [$_c$ in May 1983 $_c$] [$_d$ after trying to organize an unauthorized pilgrimage $_d$]. $_a$]

There are some properties of discourse that allow to enrich a structure like that proposed here by regular procedures, so that the computational cost of obtaining the structure is tree-like, but the final structure can be as semantically rich as a DAG. What makes this possible is the fact that some discursive relations imply others, so that the implied relations can be obtained from the explicit ones.

An illustration of one such process can be seen in Figure 3.2. Here we can see that the default structural meanings, *continuation* is transparent and allows richer meanings to percolate. Thus, if a discourse unit B establishes an *elaborative* relation with a discourse unit A, C establishes a *continuation* relation with B and D also a *continuation* relation with C, the *continuation* relations allow the *elaboration* meaning to percolate, so that it can be regularly inferred that C and D establish the same relation as B with A.

3.3 Discourse segments

A discourse segment is a chunk of text that the speaker uses to do something

Moser, Moore and Glendening (1996)

[2]If there is no overt mechanism to relate the pair, we stipulate that the marked segment is the one coming later in the sequence of text.

In our representation of discourse, discourse segments are the subset of discourse units that contribute propositional content. We are working with minimal discourse units, thus not following a very important area of work that focusses on the segmentation of discourse at higher levels, like topics or intentions.

The structure of this section is as follows. First, some previous work on discursive segmentation is presented in Section 3.3.1, highlighting the achievements and questions that remain open, like for example the relation between syntactic form, logical form and discourse units. Then, in Section 3.3.2 a definition of the concept is provided, in accordance with the framework described in the previous chapter.

3.3.1 Previous work on computational discourse segmentation

There have been a number of approaches to the concept of discourse segment. Many approaches to discourse coherence focus mainly in the relations between discourse units, and pay little (if any) attention to the nature of discourse units. It is very common among these theories to establish a one-to-one mapping between discourse segments and the units that come by default from the linguistic level immediately underlying their discursive level of analysis. This is why minimal discourse units have usually been taken to be clauses (Grimes 1975; Givón 1983; Longacre 1983; Mann and Thompson 1988; Martin 1992), sentences (Halliday and Hasan 1976) or turns of talk (Carletta *et al.* 1996).

Nevertheless, there are a number of proposals to define discourse segments, resorting both to their intrinsic and extrinsic features. Among intrinsic features of discourse segments, the most exploited are their **content**, either in terms of propositional representation or informationally, or their linguistic **form**, either ortographic, syntactic or prosodic. We will develop about these below, first let us briefly discuss the use of extrinsic features to identify discourse units.

We consider that segments are characterized extrinsically when they are defined in relation with phenomena that are beyond the scope of the segment itself. For example, Kan (2003) defines discourse units with respect to a prototypical structure of documents, in his case, medicine articles. Carletta *et al.* (1996) consider discourse units in dialogues as steps within an encompassing plan to achieve a goal, either from one of the participants in the dialogue or from both.

It can be argued that discourse units are always defined with respect to something, since they are defined as belonging to a discourse. However, plain containment relations are less constrained than the relation that is established between a content-rich structure and the elements it contains. Therefore, containment relations are less informative, but also more flexible, that is, adaptable to different representation needs and capabilities. It is not clear that all the information conveyed by a content-rich structure will be useful for all discourse representation needs.

In contrast to extrinsic properties, intrinsic features characterize discourse units with respect to properties that can be found in the unit itself. We are going to discuss two different approaches to the characterization of discourse segments by their intrinsic properties: one that defines them by their relation with boundaries at discourse level, and another that

defines them as units that convey a piece of meaning that is atomic at discursive level.

3.3.1.1 Segments delimited by their boundaries

One of the most remarkable properties of discourse units is that they are to be found between *discontinuities* in the flow of discourse (Boguraev and Neff 2000). Various kinds of discontinuities have been found indicative of boundaries at discourse level:

ortographic: as said before, the default representation of minimal discourse units is typically defined by overt formal, structural boundaries like strong punctuation in written text or turns of talk and strong pauses in oral speech (Halliday and Hasan 1976; Hobbs 1985).

syntactic or intonational: a more accurate version of the previous consists in considering the boundaries of syntactic or intonational constituents as boundaries to minimal discourse segments. Many approaches consider clauses as minimal discourse segments (Grimes 1975; Givón 1983; Longacre 1983; Mann and Thompson 1988), but also boundaries belonging to constituents smaller than clauses have been considered as marking possible discourse segments (Schauer and Hahn 2000).

It has been noted that rhetorical relationships[3] can be found even within subclausal constituents; in example (5), the relation of alternative, typically expressed between two or more clauses, is found within the scope of a noun phrase

(5) *Let's have no more of your neither-here nor-there observations.*
original example from Knott (1996)

In contrast with sentence and clause boundaries, subclausal constituent boundaries cannot be considered as discourse segment boundaries just by themselves, but further conditions have to be fulfilled, usually related to the markedness status of the boundary or to the propositional or informational status of the content conveyed by the constituent. This other kind of properties of segments will be discussed below.

Based on the view of discourse segments as clauses, Carlson and Marcu (2001) and Mimouni (2003) have devleoped systematic guidelines to increase the reliability and stability of human annotation of discourse. Their main contribution is a clear delimitation of the fuzzy concept of "clause" for the purpose of discourse segmentation. However, neither of these proposals consistently adhere to the conception of discourse segments as clauses. For example, Carlson and Marcu (2001) state that clausal subjects or objects are not to be considered as discourse segments, but some non-clausal constituents can be assigned the status of discourse segments, as long as they are marked by a discourse marker.

[3]It can be considered that rhetorical relationships take discourse segments as arguments, therefore, they can be taken as good indicators of the presence of at least one discourse segment.

There are also a number of guidelines for discourse segmentation specially targeted to annotate dialogue (Discourse Resource Initiative 1997; Cooper *et al.* 1999). These approaches rely basically on turns of talk as boundaries for discourse segments. However, they characterize discourse segments mostly by their content, relying on the notion of intention of the speaker that will be discussed below.

Hirschberg and Litman (1993) define a discourse segment as a span of speech corresponding to an intonational phrase, therefore, whenever there is a change in tone, it can be considered as a discontinuity that marks a segment boundary.

referential and intentional: Grosz and Sidner (1986)'s model of the structure of discourse has inspired many approaches to the organization of discourse, including segmentation. Grosz and Sidner propose a three-dimensional organization of discourse, where each dimension accounts for three different aspects of discourse: linguistic form and propositional representation (*linguistic level*), topics dealt with in the discourse (*attentional level*) and plans of the speakers (*intentional level*). Discourse units are defined differently in each of these levels, and so the set of units need not (and most probably will not) be corresponding across levels.

Discourse units at the linguistic level are quite comparable to the default sentence-like discourse units, both in size and in form. Attentionally and intentionally, however, discourse units are very different. For one thing, they are much better defined by their content than by their relation with boundaries. They are also much bigger than the units we have presented until now. For these reasons, we believe that these discourse units belong to a higher level of discourse, and are reluctant to call them minimal discourse segments.

Nevertheless, attentionally and intentionally defined discourse units have been very influential in work on discourse segmentation, and the notions of topic and intentions are useful to define relations between minimal discourse segments, so we are going to briefly discuss topic- and intention-based approaches to discourse segmentation.

Most knowledge-intensive studies on discourse segmentation exploit the idea that segments correspond to speakers' or writers' intentions, while data-driven approaches exploit the distribution of topics in text to characterize segments.

Passonneau and Litman (1997a) had human judges identify intention-based discourse segments in an oral narrative. Potential boundaries were boundaries of intonational phrases. They found that judges agreed significantly on which of the set of potential boundaries were actual segment boundaries. Most interestingly, they found an important correlation between marks like pauses and cue phrases and the liability of a boundary to be a segment boundary.

Passonneau and Litman established a gold standard segmentation of the corpus, and characterized segments with a number of features, including the presence of cue phrases, the nature of referential links between segments (if any), and their informational status. Then, this information was used to induce a decision tree to deter-

mine whether each potential boundary was an actual segment boundary, using C4.5 (Quinlan 1993).

Passonneau and Litman's decision tree performing best obtains a recall of 53% and a precision of 95%. Interestingly, Hearst (1994) reports that a very simple, fully automatic topic segmentation algorithm, based on the repetition of words as the only similarity measure to determine topic segment boundaries, obtains a 59% recall and 71% precision when compared to the manual topic-based segmentation of a text. Therefore, it is arguable that shallow approaches may yield satisfactory results for topic-based discourse segmentation. It remains to be seen whether such satisfactory results can also be obtained for intention-based segmentation and, more crucially for our purposes, for finer-grained segmentation.

Kozima (1993) and Hearst (1994) developed a method to characterize topic-like discourse segments automatically. Ortographic boundaries in the text (sentence or paragraph breaks) are taken as potential topic boundaries. Then, the words in each chunk of text delimited by potential topic boundaries are compared to the words of its neighbouring potential segments. If the lexic in neighbouring segments differs above a certain similarity threshold, a segment boundary is identified between the two segments, and they are said to belong to different topic segments. Similarities between words can be calculated in different ways: based on their similarity in form, including simple repetition, because they belong to a same semantic field, as encoded in a thesaurus, because they are near in an ontology like WordNet (Miller *et al.* 1991), etc. Yaari (1997) enhanced the basic method to find a hierarchical structure of topics in a text, that is, to identify progressively smaller topic segments within the ones already identified. Even though, topic-based segments are far coarser-grained than is needed to describe the local structures that are of interest to us.

It seems clear that discursive units at higher levels (topics, intentions) can be successfully characterized by their boundaries, even exploiting shallow analyses of text. We have also seen that, at lower levels, the boundaries identifying minimal discourse units are the same as the ones that characterize the biggest constituent in the linguistic level immediately preceding discourse[4], mainly sentences, clauses or turns of talk. However, we have also seen that these units fall short to provide an adequate representation of discourse, as can be seen by the fact that guidelines for manual segmentation of text go far beyond the concept of clause.

In short, we can say that the approach to defining discourse segments by their boundaries seems to provide a satisfactory representation of discourse. However, when a better understanding of the phenomenon is needed, for example, to provide guidelines to human judges to guide their task of segmenting a text, boundaries come short to define discourse segments .

We can conclude that boundaries can be a very adequate approach to characterize discourse segments in cases where only a shallow analysis of text is available, for example, for

[4]assuming a layered analysis

identifying them automatically with shallow NLP techniques. However, a more insightful definition of the concept is needed if we want to obtain a representation of discourse that is aimed to understanding the text. Although this is not feasible with state-of-the-art NLP tools, specially for languages other than English, we should bear this target in mind to guarantee consistency with future developments of the NLP technology.

3.3.1.2 Segments as atomic units of discursive meaning

In contrast to shallow analyses, knowledge-intensive approaches to discourse analysis define discourse segments by their content, rather than as chunks of text between discourse boundaries. As will be developed below, concerning their content, we consider that discourse segments are "*linguistic units conveying a piece of meaning that works atomically at discourse level*". Depending on their affinity with our approach, we distinguish two kinds of approaches to the definition of discourse segments by their content.

A very extended approach defines discourse segments as conveying a piece of content that can be treated as a unit, irrespective of the atomicity of this content. Normally, these approaches do not pay much attention to *minimal discourse units*, but are more concerned with defining the content that characaterizes them. However, it can be assumed that they consider minimal discourse units as those where no subunits can be distinguished.

A good exponent of this line of work are Grosz and Sidner (1986), who define discourse segments as linguistic units conveying a topic or intention, either simple or composed of subtopics or -intentions. This line has been very widely applied to manual discourse segmentation (Passonneau and Litman 1997a; Carletta *et al.* 1996; Discourse Resource Initiative 1997; Cooper *et al.* 1999).

The other approach that we distinguish is the one where we find ourselves in. In this approach it is assumed that a minimal discourse unit should convey the content equivalent to a proposition. The classical theory to this knowledge-intensive definition of discourse segment is provided by Polanyi (1996) under the Linguistic Discourse Model (LDM). Discourse segments are defined as a "*contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world*", or, as put simpler in (Polanyi *et al.* 2004), a unit that "*communicates information about not more than one "event", "event-type" or state of affairs in a "possible world" of some type*".

van Halteren and Teufel (2003) present a very interesting approach to determining atomic units of meaning at textual level, specially oriented to information condensation tasks. They define *factoids* as units of meaning that have to be distinguished in order to compare different summaries of the same text. The definition of the unit is very vague, because it is totally dependent on the summaries to be compared: factoids correspond to expressions in a FOPL-style semantics, compositionally interpreted. However, if a certain set of potential factoids always occurs together, this set of factoids is treated as one factoid. This supposes an important shortcoming for generalzation, because it is difficult to determine the set of factoids that represent the content for a given text. Indeed, van Halteren and Teufel (2003) shows that the number of factoids increases as new summaries are introduced to the set of

summaries for the same text, and this increment does not seem to come to a stable point as more judges are added to the initial inventory.

The approach by van Halteren and Teufel (2003) clearly points the most preocuppying limitation of an approach to discourse segments based on their meaning, namely, that meaning is a continuum that can be splitted in many different places by different judges. Therefore, formal indications of boundaries can be very useful as landmarks to assess discourse segmentation tasks. Indeed, we have seen in the previous section how discontinuities are indicative of segment boundaries. Also units[5] of linguistic form can be considered as indicative of units of discursive meaning.

It is a frequent assumption that an atomic meaning has to be expressed by an atomic linguistic form, as recently synthesized by Polanyi *et al.* (2004), who define discourse segments as "*syntactic constructions that encode a minimum unit of meaning and/or discourse function interpretable relative to a set of contexts*". However, it is not clear which "syntactic constructions" convey a minimum unit of meaning at discourse level. Polanyi *et al.* provide an extensive list of syntactic constructions that present different discursive behaviours; also Carlson and Marcu (2001) and Mimouni (2003) devote lengthy sections to describing the discursive semantics of different syntactic constructions.

We can conclude that discourse segments are usually defined by their content in well-informed theoretical models of discourse. Defining them by their content seems to provide a criterion for distinguishing them where form falls short. For example, many of the possible indicators of discourse segment boundaries are highly ambiguous, but it can be disambiguated if the content of the potential discourse segments separated by this indicator fulfill what is required for a span of text to be a segment[6]. In example (6) we can see that a comma marks a segment boundary in (a) but only an enumeration in (b), and it is the fact that the constituents separated by the comma in (a) convey the content equivalent to a full-fledged proposition that allow us to identify this comma as an effective boundary.

(6) a. [Torture of political detainees in Mauritania has been routine since 1986] , [but it has never before been used on such a scale] .

b. [Thirty-six other lawsuits against 35 National Police officers] , [three Treasury Police officers] , [10 civilians] and [three judges for a wide range of abuses against street children are reportedly pending in Guatemalan courts] .

However, content alone is also insufficient to provide a reproducible definition of discourse segment, as shown by van Halteren and Teufel (2003). In conclusion, both propositional content and formal indications of boundaries have to be taken into account to achieve a descriptively adequate definition of discourse segment that is useful for implementation in a robust discourse parser.

[5]The difference between units and boundaries is that units are endocentric, that is, that they can be seen as a hierarchical structure with a visible heade, while boundaries are properties of strings, regardless of any head.

[6]In our case, that it conveys information about a single event and that this information has a linguistic form that makes it autonomous of other information about the same event that may be conveyed in the same text

3.3.2 Definition of discourse segment

In what follows we will present our definition of a common concept of minimal discourse segment, targeted to represent relevance and coherence to improve text summarization via shallow NLP.

If various dimensions of meaning are distinguished, discourse segments could be defined based exclusively on the features that are relevant in this dimension, as is the case of Grosz and Sidner (1986). This would probably yield non-comparable segments in each dimension, which makes it more difficult to obtain a unified representation of discourse to account for general coherence and relevance. For one thing, interpretation heuristics work collaboratively with all dimensions, which makes it convenient to work with a common set of units. On the other hand, there are reasons to believe that minimal discourse segments could be thought of as a general linguistic entity, more than as a mere theoretical concept that is descriptively adequate only within a concrete model of discourse, that is, the same kind of unit as phrases or clauses. We try to support this claim with human judgements in Section 5.2.

We take *clauses* as the default minimal discourse segment[7]. The one-to-one mapping between clauses and discourse segments is oversimplifying, since most discourse segments are smaller. As we showed in Alonso and Fuentes (2002), representing text in units smaller than full sentences improves the quality of automatic summaries.

In short, we consider that discourse segments are:

- *which may be discontinuous, but in all cases self-contained,*
- *between discontinuities at discourse level,*
- *not breaking any simple syntactic or intonational constituent,*
- *not breaking apart the constituents of any argumental nucleus or prosodic curve,*
- *conveying a single event,*
- *contained by a single unit of meaning in higher levels of analysis*

Of course the exploitation of the features listed here as characterizing of discourse segments for automatic analysis is subject to the availability of automatic analyzers that can recognize argumental structure, prosodic curves, etc.

We will go deeper into the characterizing features of this definition in what follows. Since we are not particularly concerned with oral discourse, the intonational properties of discourse segments will not be discussed further.

Following previous work, our definition of discourse segment presents two different aspects: external, sufficient conditions for the identification of segments, and internal, necessary conditions for a segment to be established. We identify shallow textual correlates fro both of them, which are exploited in Section 5.4.

[7]If clauses cannot be identified within sentences, then sentences are the default discourse segment

3.3.2.1 Discourse segments are identified by their boundaries

The external evidence for identifying discourse segments are the **discontinuities** in the flow of discourse (Boguraev and Neff 2000), identified by the following shallow textual correlates:

punctuation as in the following example, where commas help to detect phrases that can function as a discourse segments.

(7) *The show, presented by Caterer Hotelkeeper, is at the Wembley Conference Centre, London, on 6 and 7 November.*

a. The show is at the Wembley Conference Centre
b. presented by Caterer Hotelkeeper
c. London
d. on 6 and 7 november

Two kinds of punctuation can be distinguished, depending on their likelihood to be discourse segment boundaries: strong punctuation, like full stops, are inambiguous evidence of boundaries, while weaker punctuation, like commas, are not reliable enough to be taken as full-fledged evidence of discursive boundaries, and need to be supported by other evidence. In the previous example, commas are supported by: a hypotactic syntactical construction in (b), a *location* phrasal head in (c) and a *time* phrasal head in (d).

syntactical structures the boundaries of syntactical constituents are potential discourse segment boundaries. Some kinds of boundaries are more likely to constitute discourse boundaries than others: the constructions headed by a verbal form, inflected or non-inflected, paratactic or hypotactic, are the most likely to constitute a discourse segment boundary, a good proof of that is that they are normally accompanied by other evidence of boundaries, like punctuation or discourse markers. In example (8), constructions (c) and (d), headed by participles, constitute discourse segments. The boundary of segment (c) is supported by a potential discourse marker , "*after*".

(8) *George Mtafu, Malawi's only neurosurgeon, was arrested after refusing to apologise for challenging public criticisms of northern Malawians made by Life-President Banda.*

a. George Mtafu was arrested
b. Malawi's only neurosurgeon
c. after refusing to apologise for challenging public criticisms of northern Malawians
d. made by Life-President Banda.

The boundaries of constructions headed by verbs are normally very marked structurally, but this is also the case for some kinds of phrases, like those conveying a big

amount of information, or those that are very periferal to the clause, as can be seen in example (9).

(9) *Over 150 men, women and children were killed by Mali's security forces in March, after a wave of pro-democracy demonstrations and riots.* extracted from BNC

a. Over 150 men, women and children were killed by Mali's security forces in March

b. , after a wave of pro-democracy demonstrations and riots.

discourse markers discourse markers can be considered discontinuities because they do not behave as content words or as clausal function words, as will be discussed in the next section. An example of a discourse segment boundary marked by a discourse marker can be seen in example (10).

(10) *I like to call it [natural gas] hemispheric in nature because it is a product that we can find in our neighborhoods.*
George Bush, declaration, Austin, Texas, December 20, 2000

a. I like to call it [natural gas] hemispheric in nature

b. because it is a product that we can find in our neighborhoods.

As we have seen up to here, different shallow clues have different reliabilities as signals of discourse segment boundaries. We have found that taking these different reliabilities into account improves the accuracy of an automatic discourse segmentation algorithm presented in Section 5.4.

3.3.2.2 Atomic meaning in discourse segments

As we have said before, in addition to boundaries, the definition of segments has to be completed by a definition of their content.

In a shallow NLP approach boundaries can be identified by simple pattern-matching operations, or exploiting shallow analyses, while systematizing content usually requires large knowledge bases and complex reasoning mechanisms.

For an illustration, let's consider the nonmarked content of a discourse segment. The argumental core of a clause can be considered as the unmarked content of a discourse segment. However, deep NLP tools are necessary to determine the argumental core of a clause. Shallow NLP falls short in this respect, because shallow correlates of predicate–argument (vs. predicate–adjunct) relations can be very ambiguous, as can be seen in example (11), where the argument of the verb (c) is signalled by the same mark, *for*, that in example (12) signals the not-subcategorized adjunct (d), liable to be a discourse segment.

(11) *Amnesty International is a worldwide human rights movement which works impartially for the release of prisoners of conscience.*

a. Amnesty International is a worldwide human rights movement
b. which works impartially
c. for the release of prisoners of conscience

(12) *The five [prisoners of conscience held in Swaziland] were previously imprisoned from June until October 1990 for allegedly organizing a political party [...]*

a. The five were previously imprisoned
b. from June
c. until October 1990
d. for allegedly organizing a political party

Therefore, the requirements on the content conveyed by discourse segments that we are providing here have to be understood more as an ideal target than as something that will be exploited in an effective system that identifies discourse segments automatically. Systematization of content is also useful to guide human judges in segmentation tasks. We expect that the development of NLP tools will make it possible to exploit these features in the near future, for the automatic identification of discourse segments.

Very intuitively, discourse segments can be defined as linguistic units conveying a piece of meaning that works atomically at discourse level. But then: what do we understand by *linguistic units*? what is an *atomic piece of discursive meaning*?

As said before, the default discourse segment is the clause. However, also other linguistic units may be considered, the only requirements with respect to their linguistic form or structure are:

`they cannot overlap with other discourse segments`

`they may be discontinious`, as in the following example, where a discourse segment is constituted by the two elements found at each side of a parenthetical segment, in this case, a relative clause, that is inserted in between.

(13) The Expert on Equatorial Guinea, a country which receives assistance under the UN Advisory Services Program, has been requested to study the human rights situation there.

`they have to be self-contained` that is, they have to be a full linguistic unit (phrase or clause), they contain exactly those elements that are necessary to constitute a complete linguistic. In the following example, the matrix clause cannot be considered an autonomous discourse segment, because the two instances of "*so*" are incomplete and require a supplementation from another constituent to constitute a complete linguistic structure, in this case, a clause.

(14) Ingredients are so fresh and so quickly prepared that food poisoning is not considered a threat.

We understand that a piece of meaning works atomically at discourse level when none of the parts in which it could be decomposed can work autonomously in its context without the resulting text being infelicitous (Alonso and Castellón 2001) or experiencing a change in the meaning it conveys, either propositionally or implicationally.

In example (15), it can be argued that (15-a) is a minimal discourse segment because it conveys a minimal unit of content at discursive level. As can be seen in (15-b), it can be decomposed in various different units that globally convey the same propositional meaning as (15-a) and are still grammatical and felicitous. However, the second phrasing conveys a different discursive meaning, mainly due to its different informational structure.

(15) a. Zog dropped the stone.
b. The stone fell. It hit the ground. Zog did it.

original example from Knott (1996)

To sum up, we believe that discourse segments are closer to *fragments* (Merchant 2003) than to clauses. Their configuration is determined by communication mechanisms, and not by the rules of clausal syntax or semantics, although it is consistent with them. The necessity for a discourse-based functional unit is discussed in Polanyi (1996), who argues that other linguistic units are not adequate for modelling discourse.

In conclusion, both external and internal features are necessary to achieve an adequate definition of discourse segment, both to guide human judgements and to shape an algorithm for the automatic identification of discourse segments, even with a shallow NLP approach.

See example (16) for an illustration. All four sentences in this example convey the same propositional information, namely:

$$\begin{aligned} &\exists BertrandRussell \quad \wedge \quad \exists book \\ &\wedge \quad write(BertrandRussell, book) \\ &\wedge \quad \neg(read(we, book)) \end{aligned}$$

However, their discursive segmentation is different. In (a), no segment can be identified because there is no constituent that works autonomously at the level of linguistic form[8]. In contrast, segments can be identified in (b) to (e), because there are sentence constituents or even sentences that can work autonomously (signalled by square brackets).

(16) a. [We didn't read Bertrand Russell's book.]
b. [We didn't read the book] [by Bertrand Russell.]
c. [We didn't read the book] [written by Bertrand Russell.]
d. [We didn't read the book] [that was written by Bertrand Russell.]
e. [We didn't read the book.] [It was written by Bertrand Russell.]

The most clear case is (e), where there are two distinct sentences, conveying two different events, so that two different discourse segments can be clearly distinguished. The rest of

[8]Note that van Halteren and Teufel would have no problem to identify a segment in (a), since they rely exclusively on the *content* conveyed by utterances, and not in their form.

the cases are more controversial. In (d), the relative clause has an inflected verb, so it is fairly objective to say that it is conveying information about a different event and should therefore be considered as a discourse segment . In (c), there is no inflected verb but the participle can be thought of as conveying an eventuality at the same level as a relative clause.

(b) is the most controversial case. Both in form and in meaning, (b) is closer to (a) than to (e): the prepositional phrase does not contain any linguistic form that could serve as the propositional basis for an event about which something is said. Yet we are willing to say that the prepositional phrase in (b) is a discourse segment , while the adjective phrase in (a) is not. There are various reasons to assert this. First, this phrase does not belong to the argumental core of the clause to which it is attached, and it is not linearly situated within it, but on the right periphery of the clause. Intonationally, this phrase has a high probability of being distinguishable within the prosodic contour of the whole sentence.

3.4 Discourse markers

In our representation of discourse, discourse markers are discourse units that contribute virtually no propositional content to the structure of discourse, but that provide information on properties of the structure itself. Discourse markers can determine the boundaries of discourse units conveying propositional content, identify relations between them and specify the structural and semantic meanings of these relations.

These functions can be carried out by many linguistic devices, but, as will be explained in Section 4.1, discourse markers are particularly well suited to our framework and objective, because they are highly informative of discourse structure, while treatable with shallow NLP techniques.

The adequacy of discourse markers as highly informative clues for formal discourse analysis is lessened by a lack of consensus about what is and what is not a discourse marker, not from a theoretical standpoint neither for applied purposes. Therefore, we will provide our own definition of the concept, a definition that is coherent with the representation of discourse we have described so far, that is sufficient for us to exploit the information provided by discourse markers for our needs and adapted to our capabilities.

This section is organized as follows. First, we give a brief overview of previous work on discourse markers from a computational perspective. Section 3.4.2 presents an indicative defintion of the concept, implemented in a lexicon in Section 3.4.3.

3.4.1 Previous computational approaches to discourse markers

Here we present a very brief overview of the computational the study of discourse markers. Since we focus in formal perspectives, we are leaving aside the extense work about discourse markers that has been carried out from the fields of discourse analysis and descriptive grammar.

As we have already pointed out, there is no consensus about what should and should not be a discourse marker. To begin with, different terms have been used to refer to linguistic elements (and sometimes also extra-linguistic: gestures, movements) with a discursive function: many different items have been named "discourse markers", and what we call discourse markers have been named by different terms, like *discourse particles, cue phrases, clue words, connectors, conjunctives, connectives* or even *textual particle*. Here we are taking the name *discourse markers* because this seems to be the standard term in the area of automatic text summarization after the work of Marcu (1997a), rooting from Schiffrin (1987).

Even if we leave apart differences in term, the scope of the concept greatly varies in different approaches[9]. Linguistic devices conveying discursive effects are not as well studied as elements whose scope can be delimited to phonological, morphological, phrasal or even clausal levels. Therefore, the distinctions between different elements at discourse level are fuzzy, and it is not uncommon that the same term (*discourse marker* or another) is used to refer to very different sets of linguistic elements, covering virtually any element whose effects require more than a propositional representation to be adequately described, ranging from words eliciting coherence relations to items that signals a relation between the participants of the communicative act.

In NLP, discourse markers have been exploited for a variety of purposes, in the hope that they would help systems to capture the intentions of the user (in the case of analysis) and to achieve more natural and comprehensible output (in the case of generation). They have played an important role in NL Generation systems (Hovy 1988; Elhadad and McKeown 1988; Elhadad and McKeown 1990; Di Eugenio *et al.* 1997) and, more recently, they have gained importance as evidence to obtain the discursive structure of text (Marcu 2000; Schilder 2002). However, most of these approaches have not bothered to provide a solid definition of the concept of discourse marker, but have just employed a set of items that have been typically considered as discourse markers (*then, because, however*) or have used sets that were defined by someone else.

In some cases, discourse markers are defined as those devices that signal the kind of discourse relation that is of interest for a given application. This approach is not useful to us, as we use discourse markers precisely to determine the set of relations to represent discourse. In other cases it has been said that discourse markers are those items that can be found at the boundaries of discourse segments. Again, we exploit discourse markers to identify discourse segments. As follows, discourse relations and discourse segments cannot be used as evidence to define discourse markers, because precisely discourse markers constitute the evidence that defines them.

Theoretical or descriptive linguistics do not provide much help with the definition of discourse marker. There are many definitions of discourse markers or similar concepts, but they are either too informal for computational use, or else they do not cover the kind of linguistic items that are useful for NLP applications. For example, the reference work in dis-

[9]For an extensive review on different perspectives about discourse markers, including a comparative exposition of different terms and the scope of these terms, see (Bordería 1998)

course markers for Spanish from descriptive grammar, Martín Zorraquino and Portolés (1999), consider as discourse markers only those items that have scope over a whole inflected sentence, like *however* or *on the other hand*, and leaving aside items like *because, but, so*, etc., which provide very useful information for NLP applications (namely, coherence relations between spans of text). The work rooting in conversation analysis defines discourse markers as proper of oral language, conveying information about the relations between (models of the) participants in the communicative act, thus leaving aside not only items like *because*, but also items like *however*.

In this context, the work of Knott (1996) is specially important because it provides a method to determine what is a discourse marker for computational purposes, shown in Figure 3.3. In two words, Knott proposes that any word or phrase that is infelicitous when it is left alone with its host clause, without any further context, must be considered a *cue phrase*[10]. The core idea of the test is that cue phrases have a supraclausal function, which is clearly seen when the clause is isolated.

Despite the importance of this work, one main criticism has to be made to Knott's work: the fact that only inter-clausal connectives qualify as discourse markers leaves aside a whole range of prepositions whose meaning is comparable to that of inter-clausal connectives, as discussed by Stede and Umbach (1998) and Schauer and Hahn (2000). Examples of the equivalence of clausal and phrasal connectives are shown below.

(17) $$\begin{aligned}&(\ irregular(count(votes))\ ) \wedge (\ win(Bush, election)\ )\\ &\quad \wedge (\\ &\qquad (\ irregular(count(votes)) \Rightarrow \neg(valid(election))\ )\\ &\qquad \wedge\ \ (\ (\neg(valid(election))) \rightarrow \neg(\forall(x), win(x, election))\ )\\ &\quad)\end{aligned}$$

a. Although votes were counted with most irregular procedures, Bush won the election.

b. Despite irregularities in the counting of votes, Bush won the election.

(18) $$\begin{aligned}&\exists(release, before(P))\\ &\quad P = (\\ &\qquad (\ arrive(x, Swaziland)\ )\\ &\qquad \wedge\ \ (\ want(x, (talk(x, government)))\ )\\ &\qquad \wedge\ \ (\ represent(x, AI)\ )\\ &\qquad \wedge\ \ (\ two(x)\ )\\ &\quad)\end{aligned}$$

a. The releases occurred shortly before two AI representatives arrived in Swaziland for talks with the government.

b. The releases occurred shortly before the arrival in Swaziland of two AI representatives for talks with the government.

[10]Knott calls *cue phrase* what we call discourse markers.

1. Isolate the phrase and its **host clause**. The host clause is the clause with which the phrase is immediately associated syntactically; for instance, if the passage of text to be examined is

   (4.1) ... John and Bill were squabbling: John was angry *because* Bill owed him money. That was how it all started ...

   then the isolated phrase and clause would be

   (4.2) *because* Bill owed him money.

2. Substitute any anaphoric or cataphoric terms in the resulting text with their antecedents, and include any elided items. For the above clause, this would result in

   (4.3) *because* Bill owed John money.

   Propositional anaphora *within the candidate phrase itself* should not be subsituted, however. Thus if the candidate phrase is *because of this*, the propositional anaphor *this* should remain.

3. If the candidate phrase is indeed a relational phrase, the resulting text should appear **incomplete**. An incomplete text is one where one or more extra clauses are needed in order for a coherent message to be framed. The phrase *because Bill owed John money* is incomplete in this sense: it requires at least one other clause in order to make a self-contained discourse. Even the fact that it could appear by itself on a scrap of paper (say as an answer to a question) does not make it complete; the question is essential context if it is to be understood.

   Note that it is only additional clausal material which is to be removed in the test. Any additional contextual information necessary for the comprehension of the clause (for instance, knowledge of the referents of definite referring expressions like *John* and *Bill*) can be assumed to be present.

4. Any phrases which refer directly to the text in which they are situated (such as *in the next section*, *as already mentioned*) are to be excluded from the class of relational phrases. Such phrases pass the test–but only because their referents have been expressly removed through the operation of the test itself.

5. Phrases which pass the test only because they include comparatives (for instance *more worryingly, most surprisingly*) are also to be excluded from the class of relational phrases. Stripped of the comparatives, such phrases do not pass the test. Comparatives like more and most introduce a very wide range of adverbials, bringing the compositional resources of the language quite strongly into play. Since we are more interested in stock words and phrases that have evolved to meet specific needs, phrases involving comparatives will not be considered as relational phrases.

6. Sometimes, more than one cue phrase can be found in the isolated clause (eg *and so*, *yet because*). In such cases, both phrases should pass the test when considered individually in the same context. In other words, the host clause should appear incomplete with either phrase.

Figure 3.3: Test proposed by Knott (1996, pg. 64) to determine whether a certain textual clue is marking a discourse relation between two clauses.

3.4.2 Definition of discourse marker

In this section we define the concept of discourse marker within our framework, shallow NLP, and with respect to our objective, obtaining a representation of discourse that is useful for relevance and coherence assessment oriented to text summarization.

In contrast with the definition of discourse segments, here we will not be exhaustive with regards to the properties of discourse markers. The purpose of this definition is to sketch the main properties of discourse markers, based in Knott's test but trying to incorporate items that do not have a clausal scope but nevertheless signal the same kind of information.

In order to complete this definition, we will provide an operative delimitation of the concept **by extension**. In the following section we will discuss how a set of prototypical discourse markers can be systematized in a lexicon. Then, in Section 5.5, we will show how discourse markers can be automatically identified by some features that are not strictly defining, in the sense that they may not necessarily be applicable to all discourse markers or else that they may be applicable to linguistic items that we do not consider discourse markers.

In short, we consider that discourse markers are:

- *occurring in written text of standard language, which can also occur in informal, oral text, but are not exclusive of this register*
- *lexical items, atomic or composed by more than one word*
- *that elicit a relation between spans of text that are discursive units, so that isolating the discourse marker and its host clause yields an infelicitous text*
- *that have scope over a clause or another linguistic unit that can be represented as a stand-alone proposition*

In this definition we explicitly leave aside all discourse markers that are proper of oral language, because our work is focussed in written text. We are thus also leaving aside the whole area of study of discourse markers that is centered in their role in dialogue. As said before, discourse markers of oral language (better known as *discourse particles*) tend to convey a qualitatively different kind of discursive information, usually eliciting relations between participants in the communicative act (for example, between beliefs, models of common ground, etc.).

We only consider lexical items, excluding punctuation and/or intonation, because the latter do not provide the kind of content-rich information that can be obtained from lexical discourse markers.

Unlike Knott, we do not only consider discourse markers that dominate a clause, but also discourse markers with comparable discursive effects, as those displayed in examples (17-b) and (18-b). We consider that discourse effects are comparable when the linguistic constituent dominated by a discourse marker can be represented as a full proposition, as is the case of (17-b) and (18-b), where the underlined spans, dominated by discourse markers, are not clauses (because their highmost syntactical projection is not an inflected

verb) but are propositionally equivalent to the underlined segments in (17-a) and (18-a), which are full-fledged clauses.

This propositionability test is coherent with the fact that discourse markers elicit a relation between spans of text with the status of discourse unit. As Polanyi (1996) states, a discourse unit is defined as a "*contextually indexed representation of information conveyed by a semiotic gesture, asserting a single state of affairs or partial state of affairs in a discourse world*". The propositionability test precisely checks that the textual spans related by a discourse marker are indeed indexed with respect to some world, that is, that they are propositions having a truth value in some possible world.

The requirement that a discourse marker relates discourse units captures the basic insight of Knott's test. Indeed, if a lexical item has a relating function, it is necessarily infelicitous when it is left alone with its host clause, without any further context.

In this definition we have not included features that have often been considered as characterizing of discourse markers: invariability, part of speech, discursive meaning... as said in the beginning of this section, this definition does not aim to be exhaustive, but only a part of a delimitation of the concept that will be completed in the following section by a description of the most characterizing features of a set of prototypical discourse markers.

3.4.3 Representing discourse markers in a lexicon

In this section we will attempt to complete the definition of discourse marker by extension. We will describe a set of prototypical discourse markers, parallel in three languages: Catalan, Spanish and English. They have been systematized to constitute a lexicon that will be used for computational purposes, more concretely, for analysis of natural language via shallow techniques to identify coherence and relevance relations useful for text summarization.

This section is structured as follows. First, we briefly expose the state of the art in discourse marker lexicography for computational purposes, showing that there are very few resources, specially few that are oriented to analysis and with a broad coverage, and none of this kind for languages other than English. Then, we describe our own lexicon, constituted by a small set of prototypical discourse markers parallel in Catalan, Spanish and English. Finally, we conclude by exposing the features that we have found most characterizing of this set of prototypical discourse markers, as represented in this computational lexicon.

3.4.3.1 Existing discourse marker lexica

Two main approaches can be distinguished in the efforts to systematize discourse markers in a lexicon: the classical lexicographic approach and those works that aim to build a resource for computational use. The first are useful to us in that they provide insightful perspectives on what is an adequate description of discourse markers, and that they aim to have a very broad coverage. However, they provide a rather informal account of the properties of discourse markers, and the set of items that they include does not necessarily correspond to our interests.

The computational approach is interesting in that the features of discourse markers are formalized and systematized, and it is rather immediate to incorporate the information in a computational resource into a computational application. However, most of the existing lexica are oriented to natural language generation, and so they describe aspects of discourse markers that are not of interest to us. Moreover, since they are to be used in systems typically working in restricted domains, their coverage is rather limited. Even more, no resource of this kind exists for Catalan or Spanish, even though there are some important lexicographic resources for discourse markers (Santos Río 2003).

To our knowledge, there is only one broad-coverage, analysis-oriented computational discourse marker lexicon, that created by Marcu (1997b) and developed in Marcu (2000).

Marcu (1997b) enhances the intial set of Knott (1996) and provides an account of more than 450 cue phrases, based on the study of 7600 text fragments. Based on an extensive corpus study, each instance of a given cue phrase is described with information that will be generalized and exploited by a constraint-based algorithm to determine the discourse relation expressed by discourse markers:

marker the discourse marker under scrutiny, all the punctuation marks that may precede or follow it and all adjacent markers, if any.

usage determines whether the discourse marker is having a sentential, discourse-semantic or discourse-pragmatic role.

position of the discourse marker under scrutiny in the textual unit to which it belongs: at the beginning, in the middle or at the end.

right boundary of the minimal unit in which the discourse marker is found.

where to link describes whether the textual unit that contains the discourse marker under scrutiny is related to a textual unit found before or after it.

rhetorical relation rhetorical relation signalled, one of a set of RST-like relations[11]

rhetorical statuses of the related spans (nucleus or satellite, following RST).

textual types of the related spans (clauses, sentences, paragraphs, multi-paragraphs).

clause distance between the spans related by the discourse marker, in clause-like units.

sentence distance between the spans related by the discourse marker (sentences).

distance to salient unit to the textual unit that is the most salient unit of the span that is rhetorically related to a unit before or after that under scrutiny (clauses).

Stede and Umbach (1998) discuss how discourse markers should be represented in a lexicon for computational purposes, both generation and analysis, although they do not describe the impelementation of their proposal. They argue that discourse markers have properties of open-class words and of closed-class words, like near-synonymy and hyponymy relations, and that both should be reflected in their representation. The actual work of developing the discourse marker lexicon is oriented to German, but the discussion of the features is aimed to be language independent. The properties that they find relevant for their computational usage are:

[11] In the case of polysemic discourse markers, two different entries are created.

relatedness in meaning they claim that, like typical open-class words, discourse markers can establish meaning-based relations among them, like *synonymy, plesionymy, antonomy, hyponomy*, etc.
part of speech
position in the constituent (clause or phrase)
linear order in case of co-occurrence with other discourse markers
negation special behaviour towards negative polarity
semantic relation they express: cause, adversative, conditional, etc.
effects on polarity whether they affect the polarity of the units under their scope
commentability whether they can be taken as propositional content
intention associated to the discourse marker
associated presuppositions
associated illocutions
stylistic constraints
discourse relation (different from semantic relation)

It can be seen that, even if these two approaches are both computational and, in principle, both oriented to natural language analysis, they are significantly different with respect to the features by which discourse markers are described. This difference seems to be rooted in the fact that Stede and Umbach seem to rely in a deep analysis of text, including an account of intentions and presuppositions, while Marcu is explicitly based on a syntactic analysis only. Note also that, while Stede and Umbach aim to describe *types* of discourse markers, Marcu account for *tokens*, paying little attention to the process of generalizing information from particular instances.

Even despite differences, both seem to share a common core: the semantics of the discourse relation signalled by the discourse marker. This seems to be the most crucial aspect of such a lexicon, and we will devote next chapter to it. Other comparable features are also related to the meaning conveyed by discourse markers. For example, the position in the constituent seems to be useful to specify the meaning of some kinds of ambiguous discourse markers. This feature is strongly related with the co-occurrence with certain patterns of punctuation. Also the co-occurrence with other discourse markers seems to be useful to determine specific meanings.

Also in both approaches, discourse markers are characterized by a set of features that specify how they should be treated algorithmically in order to obtain an adequate analysis. We believe that this kind of information does not belong to a static lexicon, but has to be integrated in the corresponding procedures. Most of the features described by Marcu (1997b) are algorithmic: they are oriented to determine the boundaries of the segment introduced by a given discourse marker (right boundary, textual spans) or the attachment point of this segment in the structure of discourse.

The two approaches mainly differ in the rest of the information that is provided. Stede and Umbach (1998) aim to provide a detailed description of the information state of the participants in the communicative act, which is useful for natural language generation

but not necessarily for analysis. They describe many features that rely on a deep semantic representation of texts, like effects on polarity. Then, they also consider contextual constraints, like style, which may be useful for an analyzer that treats a very broad range of texts. It is evident that most, if not all of these features are well beyond the scope of shallow NLP techniques.

In contrast, Marcu (1997b) characterizes discourse markers by surface features of their context of occurrence, which are treatable by shallow techniques. These features do not concern the semantics of the discourse marker, but rather their effects in the configuration of the structure of discourse. Features like *distance to...* contribute to determine the attachment point, and the feature of *right boundary* determines the boundaries of the discourse segment where a given discourse marker is embedded. We believe that this kind of information is not proper of discourse markers, but of the discursive structure and of the discursive segments, respectively, and that it should be treated in the correspoonding modules.

In our implementation, we have treated these two kinds of information differently. On the one hand, a discourse marker lexicon encodes all the information intrinsic to discourse markers as isolated lexical items, that is, their semantics and part of speech. On the other hand, the effects of discourse markers with respect to the configuration of the structure of discourse or discourse segments is exploited by specific procedures (see Appendix B and Section 5.4 for specifications of these procedures). These procedures rely on discourse markers as a source of evidence that is external to their processing core. It is true that discourse markers may significantly improve the performance of these procedures, as shown in Section 5.4 and also by Reitter (2003b), but the backbone of the procedure is purely heuristic.

One last and evident conclusion that we draw from this very brief review is that discourse marker lexica are a scarce resource, although they are crucial for discourse-based NLP. One of the reasons of this scarcity may be that they are very costly to build and port to different applications. As a direct consequence, most of the natural languages do not have such resource available.

In the following section we present a lexicon of discourse markers in Catalan, Spanish and English. It has been created by an economic method that significantly reduces the effort of development, mainly by characterizing discourse markers by purely semantic features, leaving all algorithmic characterization for other modules of processing.

3.4.3.2 A trilingual lexicon of prototypical discourse markers

In this section we present the lexicon that is displayed in Appendix A. There, the prototypical discourse markers that constitute this lexicon are characterized by the semantics of the discourse relations they elicit, described in the following chapter. Interesting particularities of particular discourse markers are also discussed there.

We have constituted this set of prototypical markers selecting those discourse markers from existing computational lexica (Knott 1996; Marcu 1997b) for which a near-synonym could be found in Catalan, Spanish and English, and which were above a certain degree

	revision	cause	equality	context	**total**
elaboration	4	9	10	22	41
continuation	9	9	6	4	28
underspecified	1	–	10	4	15
total	14	18	26	32	84

Table 3.1: Distribution of the number of discourse markers accross the different meanings. Some discourse markers have been assigned to more or less than one meaning per dimension, because they are ambiguous or underspecified, respectively.

of grammaticalization (Section 4.2.2.2 describes how the degree of grammaticalization is obtained). These two constraints guaranteed that these were discourse markers conveying very basic discourse meanings. Moreover, we chose a set of a manageable size, in order to study each discourse marker in depth. Since we wanted the set to be representative, some discourse markers that met the above constrains were discarded, if the meaning they conveyed was sufficiently represented.

All in all, the lexicon is formed by 84 discourse markers, representing different discursive meanings as can be seen in Table 3.1. As explained in the previous section, discourse markers are characterized only by their intrinsic properties as isolated lexical items: the meaning of the discourse relation they may convey and their part of speech.

With respect to semantics, each discourse marker has been characterized by one of the possible values in each of the two dimensions of the meaning of discourse relations, as presented in Section 4.2.3. It has also been stated whether a particular discourse marker is ambiguous between two possible values in any of the dimensions, or whether it is underspecified with respect to the meaning described in a dimension. In practice, ambiguity and underspecification are treated the same: no information is provided for that dimension.

With respect to the part of speech, we only make a distinction between three main morphosyntactic classes of discourse markers: *conjunctive*, *adverbial* and *phrasal*, the latter grouping together prepositions and subordinating conjunctions. This classification was determined empirically, based on some experiments on clustering discourse markers (Alonso *et al.* 2002a; Alonso *et al.* 2002b). In these experiments, instances of a preliminar discourse marker lexicon containing 577 Spanish discourse markers (including coordinating conjunctions and some syntactical structures, like relative clauses) were described by a set of 19 features obtained by shallow analysis of the context of occurrence of discourse markers in text. Then, these instances were clustered by their similarity according to these features.

Figure 3.4 displays the organization of the features characterizing discourse markers, organized hierarchically by their discriminating power. It can be seen that the syntactical properties of discourse markers were most powerful to cluster together homogeneous discourse markers. The main distinction was made between extra-sentential (*adverbial*) and intra-sentential discourse markers. The latter were subdivided into rightwards directed and bi-directional; this last group was formed by coordinating conjunctions, which are not included in our final lexicon because they are highly ambiguous. Three groups were

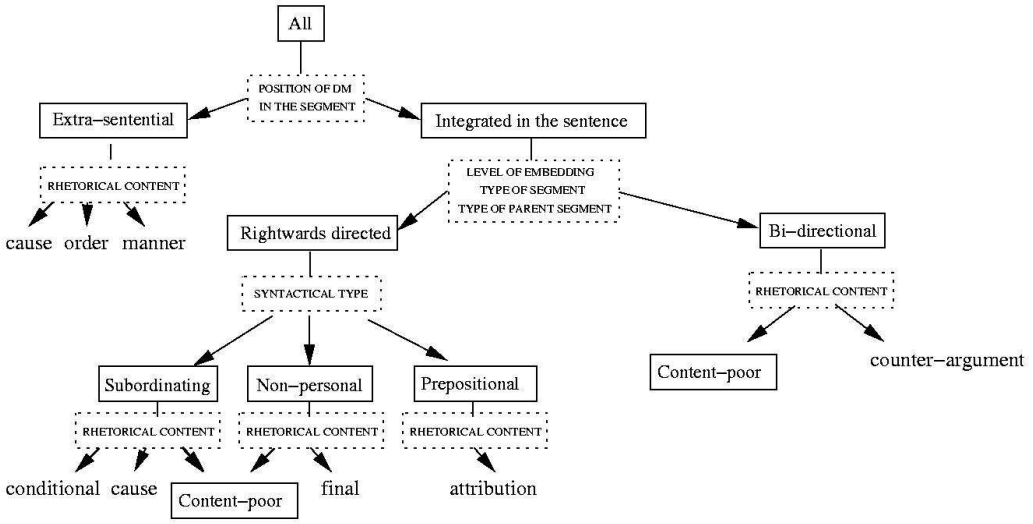


Figure 3.4: Features characterizing discourse markers, organized hierarchically by their discriminating power in clustering.

found within rightwards directed: syntactical structures (like relative clauses), prepositional (*phrasal*) and subordinating (*conjunctive*) discourse markers.

The two main groups of discourse markers are distinguished by their distribution in the sentence: *phrasal* and *conjunctive* discourse markers occur at the beginning of the linguistic constituent they dominate (phrase, clause, or fragment), with strong restrictions as to the elements that can precede them (typically, only adverbs, as in (19)). In contrast, the occurrence of adverbial discourse markers is less constrained (20).

(19) "A picture is worth a thousand words" precisely **because** it is a part of a complex, usually unconscious, web of references.

(20) a. **However**, other prominent prisoners of conscience remain behind bars.
b. Most of the reported deaths, **however**, were torture in both military barracks and police stations.
c. They don't feed when the water is cold **however**.

Conjunctives are the discourse markers with strongest distributional restrictions. Just like phrasals, they always occur at the beginning of the linguistic constituent they dominate, but they cannot be preceded by any linguistic item (21-a), and they must be necessarily preceded by the constituent to which they are related (21-b).

(21)

i They are held **because** they opposed the government's policy towards the Shi'a community.

ii They opposed the government's policy towards the Shi'a community **so** they

are held.

a.

i They are held *precisely* [because] they opposed the government's policy towards the Shi'a community.

ii # They opposed the government's policy towards the Shi'a community *precisely* [so] they are held.

b.

i [Because] they opposed the government's policy towards the Shi'a community, they are held.

ii # [So] they are held, they opposed the government's policy towards the Shi'a community

Adverbial and phrasal discourse markers do not only differ with respect to their distribution, it has often been argued that they also differ with respect to their semantic properties. More concretely, Cresswell *et al.* (2002) and Webber *et al.* (2003), among others, have claimed that adverbial discourse markers have anaphoric properties. However, as we have already explained, anaphoric properties cannot be properly exploited by shallow techniques, and we are therefore disregarding these properties in our account.

The lexicon presented here is a resource that can be used in many NLP applications. It is an important module for the discourse segmenter and interpreter presented in Chapter 5, which have been exploited in text summarization in Alonso *et al.* (2004b). The information it contains is immediately usable by shallow approaches, but is also useful for deeper approaches.

Since this lexicon is a crucial resource for these applications, any improvement in the resource will likely improve the performance of the applications. In Section 5.4 we assess the improvement in recall that can be obtained in the automatic identification of discourse markers by enhancing the starting lexicon, adding those discourse markers that can be found in the text. Other improvements that could be of much use for computational applications include:

- associating discourse markers with an index of reliability about their discursive (vs. sentential) function, which would allow to increase the accuracy in the disambiguation of their actual function in a given text.
- associating discourse markers to different heuristics to disambiguate their semantics.
- determining the semantic and functional value of co-occurrences of different kinds of discourse markers, and, in general, a more extensive encoding of the interaction between the context and the semantics of discourse markers (as in Marcu 1997b).

In sum, in this section we have presented a delimitation of the concept of discourse marker that is useful for our representation purposes and the capabilities of our framework. We have discussed how the concept has been treated in other computational approaches and have proposed a working definition of the concept that has been applied to a set of prototypical discourse markers.

We have argued that for computational applications, discourse markers are best represented in an autonomous data structure that can be used as a module by different applications, that is, a lexicon. We have argued that, in a computational lexicon, discourse markers should be characterized by intrinsic properties (their semantics and morphosyntactic class), leaving procedural aspects for the various algorithms that will exploit them. Following this criterion, we have created a lexicon of prototypical discourse markers, parallel in three languages: Catalan, Spanish and English. This lexicon is the main source of discursive knowledge for our discourse analysis implementations (Section 2.4.2.1, 2.4.2.2 5.4). It has also been the basis of a bootstrapping approach to create a bigger lexicon of discourse markers applying lexical acquisition techniques (Section 5.5).

3.5 Discussion

In this chapter we have specified a computational representation of discourse tailored to the capabilities of shallow NLP techniques and the needs of automatic text summarization. We have discussed the limitations of our targeted representation of discourse, concluding that we cannot achieve the encompassing, homogeneous structures proposed in foundational works about computational discourse analysis.

We have argued that, given our NLP capabilities, representing discourse as a sequence of local structures is reliable and allows to capture basic discourse relations whereupon judges show an important degree of agreement when summarizing texts (see Section 5.2 for details on the empirical support for this claim). We have also proposed to represent different kinds of meaning independently, in order to implement underspecification. These proposals have been formalized in a multidimensional representation of discourse, where each dimension is captured by a strict tree-like structure. This representation seems to be able to capture some configurations of discourse relations that that could only be represented by crossing branches in simple tree-like representations of discourse. After formalizing the discourse structure to implement our representation, we have distinguished two kinds of discourse units: those that provide mainly propositional content (*discourse segments*) and those that provide procedural content about the boundaries of discourse segments and/or about the relations holding between them (*discourse markers*). We have discussed theoretical and applied work concerning these two concepts, and have proposed working definitions for both. In the case of discourse markers, we have also developed a resource, a computational lexicon of prototypical discourse markers, parallel in Catalan, Spanish and English, where we have implemented the definition of the concept.

However formal it may or may not be, what we have proposed in this chapter is no more than a convenient stipulation to achieve the targeted representation of discourse based in our NLP capabilities. In Chapter 5 we attempt to support these theoretical claims with two kinds of empirical support: the interpretation of naive judges about the structure of discourse and the automatic analysis of some aspects of the proposed representation of discourse.

Chapter 4

Meaning in discourse relations

What can be said at all can be said clearly, and what we cannot talk about we must pass over in silence

Ludwig Wittgenstein, introduction to the Tractatus Logico-Philosophicus (1922)

In the previous chapter we have specified a computational structure to represent relevance and coherence relations by shallow NLP. In this chapter we describe the meaning that is conveyed by the discourse relations between segments, focusing on relations that can be obtained by shallow NLP. We claim that, despite the limitations imposed by shallow NLP, this representation is descriptively adequate, theoretically sound and empirically motivated, because it relies on linguistic phenomena that can be formally recognized and systematized.

First, in Section 4.1, we describe the shallow textual clues that we can exploit to identify discourse relations and their semantics, treatable by shallow NLP and useful for text summarization. These clues are are discourse markers, partial syntactical structures and punctuation; a set of phenomena that are immediately beyond clausal scope, whose meaning and scope can be reliably determined by shallow techniques, and which elicit coherence and relevance relations between discourse units.

The scope and meaning of the discourse relations that we will work with is delimited by way of these clues. We discuss their reliability to obtain a representation of discourse, and we find that, with shallow NLP techniques, these clues only provide reliable information about the local configuration of discourse, that is, about intra-sentential and inter-sentential relations. Therefore, we will represent the discursive organization of a text as a concatenation of such local structures.

Concerning meaning, we argue that the meaning of discourse relations is best described compositionally. In Section 4.2.1 we discuss how a compositional approach allows to deal with the ambiguity of these clues via underspecification. Then, we describe our method to determine an inventory of basic discourse meanings in Section 4.2.2. This method takes advantage of our compositional approach to determine the inventory of meanings combining *a priori* representation needs and empirical data.

For our purposes (identifying relevance and coherence relations) two main kinds (or dimensions) of discursive meaning have been distinguish *a priori*: structural and semantic. The meaning of each of these dimensions has been discretized in two and five distinct discursive meanings, motivated by the evidence provided by discourse markers and structural economy. In Section 4.2.3 we describe each of these meanings, relating them with discursive meanings that have been widely used in previous work.

4.1 Linguistic phenomena with discursive meaning

In this section we describe the linguistic phenomena that we will exploit to characterize a representation of discourse that is useful for text summarization.

There are quite a number of linguistic phenomena that contribute to configure the discursive level of texts. Some of the linguistic phenomena that are known to yield effects at discursive levels are: some syntactic structures (Scott and de Souza 1990), the patterns of anaphoric expressions (Grosz and Sidner 1986), all kinds of co-reference (including those that require complex reasoning processes, like bridging anaphora), patterns of tense and aspect (Lascarides and Asher 1993), information structure, lexical relations based on semantic fields (synonymy, antonymy, part-of) (Morris and Hirst 1991), the organization of genres (narratives, argumentation, technical discourse) (Teufel and Moens 1998), etc. This list is non-exhaustive, the configuration of the discursive level of texts is still far from well-known. What is more, many of the discursive phenomena that are known are described intuitively, and from those phenomena whose effects have been systematized, only some can be exploited with shallow NLP techniques with a certain degree of reliability.

And even more: discourse relations can also hold without the presence of any of the above mentioned phenomena. In example (1) a reader can identify a causal relation between the two sentences in all the cases, irrespective of the textual clues that mark it, if any. In (1-d) we cannot identify any textual clue that provides evidence of the relation, but still a causal interpretation is possible. In this case, a reader can establish the causal relation between the two sentences if the adequate context is provided, relying on world knowledge and reasoning abilities[1]. For our purposes, we will only study those relations that are realized by textual clues that are formally identifiable and systematizable, as is the case of (1-a) to (1-c), and we will restrict ourselves to those that can be treated by shallow NLP.

(1) a. Eva can speak Dutch because she has been studying it for 3 years.
b. Eva can speak Dutch. She has been studying it for 3 years.
c. Eva can speak Dutch. Eva has been studying Dutch for 3 years.
d. Eva can speak Dutch. Eva studies Dutch.

Some discursive phenomena, like reported speech or discourse markers, have been systematized for computational use by different approaches, but there is no consensus as to how

[1]These abilities allow us to infer that if one can speak a language and one studies it, the language is not the native language of the person and the person can speak it because she has been studying it.

they are best characterized.

Most of these linguistic phenomena are beyond the scope of shallow NLP, at least for Catalan and Spanish: for example, a good account of information structure requires full parsing and a representation of the functional layers of clauses, together with word order; none of these is yet satisfactorily solved for Spanish or Catalan. As for the thematic structure of texts, that is, their organization in topics, subtopics and comments on these topics, various recent approaches have successfully represented it using shallow approaches, but with a the underlying assumption that the meaning of words is not ambiguous. However, words are indeed ambiguous, disambiguating them is a problem still to be solved, currently far beyond the capabilities of shallow NLP. Therefore, we have to consider that approaches like the structuring of text in topics or by lexical chains is essentially not within the capabilities of shallow NLP.

In contrast, there are some linguistic phenomena whose meaning and scope can be reliably determined by current shallow techniques, and which elicit coherence and relevance relations between discourse units. These are linguistic phenomena immediately beyond clausal scope: discourse markers, partial syntactical structures and punctuation. We will describe their properties for automated discourse analysis in what follows.

4.1.1 Shallow linguistic phenomena with discursive meaning

The set of discursive phenomena that are treatable by shallow NLP techniques has changed with the evolution of the field. For example, significant improvements have been achieved to recognize the thematic development of texts (Kozima 1993; Hearst 1994).

Of all linguistic phenomena with discursive effects, discourse markers are specially useful to obtain the kind of representation we are aiming for. On the one hand, they convey rich discursive meanings but are treatable by very simple techniques, like pattern-matching. On the other hand, the kind of meanings they convey are very useful for text summarization, because they serve to identify coherence and relevance relations in text. As a disadvantage, discourse markers are very ambiguous with respect to their function or the meaning they convey. The following examples illustrate this.

In (2), *so* can be easily identified by pattern-matching techniques, but it is ambiguous with respect to its function: in (2-a) it has a sentential function, modifying an adjective, while in (2-b) it has a discursive function, linking two clauses with a causal relation. In contrast, in (3), the discourse marker *but* is unambiguously performing a discursive function, but it is ambiguous with respect to the relation it signals: in its first occurrence it signals a contrast, while in the second it serves to introduce a new topic in the text. The kind of ambiguity of *por eso* (*for that reason*) in (4) is still of a different kind: in (4-a), *por eso* signals an intra-sentential relation, while in (4-b) it relates the clause where it is found and the first clause in the text.

(2) a. Nowadays, impartiality need not be expressed quite so crudely.
b. [He] was then transferred to Marrakech where his family lives, and so could visit him regularly and provide the food necessary for his diabetic diet.

(3) It will be seen, therefore, that whatever the state of the pound Mrs Thatcher is in no danger from the constituencies. But the menace of the IRA is everywhere. Mrs Thatcher's car is armour-plated, the platform party leaves in a tank-like bus, the town crawls with policemen, and all this vigilance costs 1.1m pound. The ratepayers are sick of paying 49 of this bill.
But let me get to the various leaders' speeches.

(4) a. En España nadie escucha a nadie, por eso se produce ese inmenso griterío.
b. El ciudadano vasco Jesús María Pedrosa Urquiza jamás podrá inscribirse en el manicomial censo electoral de libre adhesión que Euskal Herritarrok propugna crear en la futura Euskal Herria independiente. Pese a sus apellidos, Pedrosa Urquiza no debía de ser un "buen vasco" y seguro que no combatía los matrimonios con "gentes de extrañas razas", que es como los sabinianos nacionalistas definían a los emigrantes, a quienes ya hace cien años propugnaban "aislar y tratarles como a extranjeros, hasta en la amistad y el trato" en cuanto Euskadi lograra la independencia. Pero hete aquí que Pedrosa Urquiza iba a convertirse dentro de un mes en consuegro del también ciudadano vasco Juan María Bizkarraga, que no oculta su condición de simpatizante del PNV. Es más, el ciudadano Urquiza era un viejo afiliado de ELA-STV, el sindicato autóctono que el primigenio delirio sabiniano pretendía convertir en una organización que aglutinara exclusivamente a los "obreros euskerianos excluyendo a los maketos". No satisfecho con todo ello, el ciudadano vasco Urquiza desoía a su partido y cada día se tomaba sus potes y vermuts en el "batzoki" del PNV de Durango. Tamaño ejercicio de ciudadanía no podía quedar impune. Quizás por eso le han matado; porque el ciudadano Urquiza era además del PP.

Considering that discourse markers are the most informative clues at a shallow level, we have to tackle these three kinds of ambiguity to exploit their discursive information. In the rest of this section we will deal with the ambiguity with respect to their discursive or sentential function. In Section 4.1.2 we deal with the reliability of shallow phenomena to obtain a representation of discourse, including the question of the ambiguous scope of discourse markers. The ambiguity with respect to the kind of meaning conveyed by discourse markers is handled in section 4.2.

The ambiguity of discourse markers with respect to their discursive or sentential function has been addressed by manually created classification algorithms (Hirschberg and Litman 1993) and by machine learning techniques Litman (1996). Studies on discourse segmentation (Passonneau and Litman 1997b) also try to determine when a certain clue is actually signalling a discursive boundary. Different kinds of evidence have been exploited to determine the sentential or discursive usage of these clues, with a special focus on prosody (in oral texts) or punctuation (in written texts).

In our approach, we will work with written text, where punctuation is a major source of evidence. Punctuation can perform a discursive function by itself, mostly to determine segment boundaries, but it is highly ambiguous in that function. In contrast, it is very useful to disambiguate the discursive or sentential function of discourse markers, as will be discussed in Section 5.4.

Moreover, punctuation and partial syntactic structures (chunks) can be sources of evidence for discourse structure also by themselves, as in example (5), where the segment *a 43-year-old area sugar company manager* can be identified as a segment because it matches the pattern `beginning-sentece noun-phrase comma noun-phrase comma verbal-phrase`. The kind of syntactical information that we exploit are untensed verbs in absolutive constructions, subordinating structures (relatives, subordinating conjunctions) and some patterns of chunks and punctuation (as `beginning-sentece noun-phrase comma noun-phrase comma verbal phrase`).

(5) Thoza Khonje , a 43-year-old area sugar company manager, was arrested on 28 February 1989.

Interestingly, discourse mechanisms can interact with propositional structure, contributing to enrich an underspecified representation of propositional content: ambiguous attachment of constituents (VP, NP or clause), argument-adjunct distinctions (light verbs, time and space, manner, cause), etc. Therefore, the propositional (semantic and syntactic) interpretation of the sentence can be left underspecified, which is precisely the state of the art of Spanish and Catalan NLP, as well as other languages with less resources. Moreover, it is clear that some properties of the propositional representation of clauses have to be solved at discourse level. This is the case for the reference of pronouns and definite descriptions, but it can also be the case for some kinds of adjunct-argument distinctions, for the optimal realization of arguments, depending on the constraints imposed by each language.

Other kinds of linguistic phenomena that have also been found useful to obtain a representation of discourse are the so-called *lexical chains*. Lexical chains try to capture cohesion relations, assuming that they are represented by semantically related items in a text. Indeed, in Alonso and Fuentes (2002) and Alonso and Fuentes (2003) we showed that automatic summaries can be improved if they are based in a shallow approach to the represenation of discourse that combines lexical chains with the above mentioned evidence (chunks, punctuation and discourse markers), in contrast with summaries obtained by any of these two approaches alone.

In this thesis we will not integrate the kind of information provided by lexical chains to configure a representation of discourse, because they capture aspects of texts that are qualitatively different from the local structures that can be obtained with the rest of evidence presented so far. Moreover, as explained above, even if lexical chains can be obtained by shallow techniques, their underlying assumptions require a knowledge-rich analysis of texts. Again, this is not the case for the evidence presented so far: usually, the kind of information provided by discourse markers or syntactical structures is self-contained, and richer or finer-grained representations are obtained compositionally from the interaction of the information provided by these devices with other kinds of linguistic information, regardless of the techniques used to obtain it.

To sum up, the linguistic phenomena that we are going to exploit to obtain our targeted representation of discourse are partial syntactical structures, product of shallow parsing, punctuation, and, above all, discourse markers, because these are the most adequate to our

NLP capabilities and they provide information on the coherence and relevance relations between discourse units. Moreover, we feel that the account of these phenomena that we are providing now will be compatible with enriched representations of discourse, obtained with more powerful NLP tools (for other languages, or in future developments).

These phenomena may provide information about the structure of discourse at very different levels, but this information is only reliable at a local level, as will be discussed in the following section.

4.1.2 Reliability of shallow cues to obtain discourse structures

As illustrated by example (4), the shallow linguistic phenomena described above may provide information about the organization of discourse at different levels. This can be considered as an ambiguity in the scope of discourse markers. If discourse markers are considered as argument-taking operators (Forbes *et al.* 2002), some like *however, for this reason* or *in contrast* (the so-called adverbials by Forbes *et al.*) are ambiguous with respect to the syntactical and/or semantic type of arguments they may take.

This ambiguity in scope is harder to solve for long-distance discourse relations than that of local relations. In human annotation of discursive corpora, the disagreement among judges increases for long-distance relations, as reported by Forbes *et al.* (2002). We find comparable results in Section 5.3. It has also been argued (Knott *et al.* 2001) that the kind of relations that hold at local level is qualitatively different from those at higher discursive levels.

Therefore, we will only work with information about the local configurations of discourse, that is, about intra-sentential and inter-sentential relations. We will not obtain a single structure that covers all the text, but a set of unconnected structures that represent different spans of text. As we have just said, this representation seems to be descriptively adequate, and it is reliable enough to support robust NLP applications.

To increase the reliability of the information we obtain from shallow cues while minimizing information loss, we distinguish different kinds of meanings that may be conveyed by the cues we exploit, as explained in the following section.

4.2 Discursive meaning inferrable from shallow linguistic phenomena

In this section we inspect the kind of discursive meaning that can be obtained by exploiting the shallow linguistic phenomena described in the previous section, focusing in their utility to identify relevance and coherence relations and in their reliability when treated by shallow NLP techniques.

As explained above, shallow clues are highly ambiguous with respect to their meaning. In addition to the kind of discursive meaning that they convey, we want to determine which of this meaning can be obtained with shallow NLP techniques with a reliable degree

of certainty. We attempt to increase the reliablity of the meaning obtained from these clues in two main ways: we delimit the kind of meaning to meet the restrictions of shallow NLP, and we distinguish heterogeneous kinds of meaning.

The advantages of distinguishing heterogeneous kinds of meaning are explained in Section 4.2.1. Partitioning meaning implies partitioning ambiguity, so that the ambiguity of a certain discursive clue can be reduced to only part of its meaning, while the rest of the meaning can be asserted with a reliable degree of certainty. Describing discursive phenomena and discourse relations as a conglomerate of distinct meanings allows to capture certain phenomena that have posed problems for the description of discourse relations. In general, this compositional description is more transparent and flexible than other approaches.

In Section 4.2.2 we propose a methodology to determine an inventory of discourse meanings that is adequate to describe coherence and relevance relations by shallow NLP. This is a basic question for all theories of discourse, regardless of how discursive meaning is described (compositionally or by atomic labels). In many cases, this inventory is totally shaped by application needs, as in most natural language generation systems. In other cases, researchers have tried to come up with a methodology to induce a set of relations from linguistic or psychological data, in order to avoid *ad-hoc* stipulations. We take advantage of our compositional approach to the semantics of discourse to combine the advantages of these two approaches: we determine the kind of meaning that is of interest to us *a priori*, and then we motivate the discretizations in these meanings based on empirical data. We discuss what qualifies as empirical evidence, and we come to the conclusion that basic discursive meanings can be inferred by highly grammaticalized discourse markers and their cross-linguistic patterns of meaning, as captured by semantic maps.

Then, in Section 4.2.3 we apply this methodology to determine the inventory of discursive meanings to describe the semantics of discourse phenomena and, consequently, of the discourse relations by which we will represent discourse structure. First, we identify the two main kinds of discursive meaning that are useful for our representation purposes (identifying relevance and coherence relations in text): the structural and semantic dimensions of meaning. Then, we make further distinctions within these two dimensions, obtaining an inventory of meanings that are finally used as features for the compositional description of discourse relations.

4.2.1 Advantages of compositional discourse semantics

In this section we argue that describing discourse relations as a conglomerate of distinct discursive meanings is more economic and transparent than with the traditional atomic labels, while expressivity is maintained or even increased. Moreover, distinguishing heterogeneous meanings can contribute to to increase the reliablity of the meaning obtained from shallow textual clues, because ambiguity can be reduced to only one part of their semantics.

A very common approach to describing the semantics of discourse relations consists in defining a set of atomic labels and associating each relation to one of them (Hobbs 1985; Mann and Thompson 1988). It has often been noted (Hovy and Maier 1995; Knott 1996)

that such a relation-based approach is problematic. The most common problems are that the inventory of relations is very big, that it is asystematic and that it fails to provide a satisfactory description of some discourse relations.

In contrast, describing discursive phenomena and discourse relations as a conglomerate of distinct meanings is more transparent, economic and flexible than relation-based approaches. Moreover, a compositional description of discourse relations allows to capture certain phenomena that have posed problems for the description of discourse relations, as we explain in what follows.

4.2.1.1 Descriptive adequacy of a compositional description

There are two main reasons why a compositional description of discourse relations is more adequate than associating each relation with an atomic label. First, some discourse relations cannot be properly described by atomic labels. Second, the expressivity of a compositional description of discourse relations is equal, if not superior, to that of its equivalents in terms of atomic relations.

Concerning the first aspect, it has often been argued that a multi-dimensional analysis is necessary to account properly for the structure of discourse. With the classical example in (6), Moore and Pollack (1992) argue that segments can establish two of the relations proposed by RST (Mann and Thompson 1988), namely an Evidence relation and at the same time a relation of Volitional Cause.

(6) a. George Bush supports big business.
b. He's sure to veto House bill 1711.

(Moore and Pollack 1992, pp. 539-540)

Also the kind of phenomena shown in the example of Figure 3.1 has posed problems to be represented by atomic labels in a tree-like representation of discourse, because, in some cases, the relations signalled by different discourse markers have to be expressed as crossing branches in a tree-like structure. As discussed in Section 3.2, some researchers have proposed that some relations are not tree-like, but anaphoric. Instead, we propose that a distinction between heterogeneous kinds of meaning in co-occurring dimensions of meaning also provides a satisfactory description of such phenomena.

Concerning the expressivity of compositional descriptions, for any description of discourse relations that is based in atomic labels, a trivial compositional equivalent can be found, as exemplified in Table 4.1. Therefore, the expressivity of a compositional approach is at least equal to that of an approach based in atomic labels; in both cases expressivity is crucially determined by the inventory of meanings whereby relations are described.

However, if the transformation of atomic labels is non-trivial, the expressivity of the description formalism can be increased without increasing the size of the inventory of meanings that are used to describe discourse relations. An example can be seen in Table 4.2, where a set of 8 basic meanings organized in 3 dimensions allows to distinguish 12 relations. A more detailed description of this proposal to describe the semantics of discourse relations can be found in Alonso *et al.* (2003d). The organization of meanings in dimensions restricts

atomic label	consequence	purpose	reason
consequence	yes	–	–
purpose	–	yes	–
reason	–	–	yes

Table 4.1: Trivial transformation of the semantics of some atomic labels to a compositional description.

atomic label	dimension 1	dimension 2	dimension 3
consequence	cause	continuation	symmetric
purpose	cause	continuation	asymmetric
reason	cause	elaboration	asymmetric
sequence	parallel	continuation	symmetric
summary	parallel	continuation	asymmetric
example	parallel	elaboration	asymmetric
contrast	revision	continuation	symmetric
counterargument	revision	continuation	asymmetric
concession	revision	elaboration	asymmetric
narration	context	continuation	symmetric
conclusion	context	continuation	asymmetric
circumstance	context	elaboration	asymmetric

Table 4.2: Description of the semantics of atomic labels for discourse relations as a conglomerate of more basic discursive meanings, organized in three dimensions to restrict possible combinations.

their possible combinations, so that meanings belonging to the same dimension are never combined, as in Halliday and Hasan (1976).

Even if the combinations of meanings are restricted (by dimensions or any other means), the increase in expressivity can provide more configurations of meaning than are really needed to account for naturally occurring phenomena. For example, in Table 4.2 not all possible combinations of meanings occur: there is no relation characterized by *elaboration* and *symmetric* at the same time.

This excessive expressivity can affect the performance of natural language generation systems or it can suppose an unnecessary burden for analysis. As will be explained in the next section, the analytical perspective provided by a compostional approach can be of much help to design an economic system of discursive meaning, where most of the possible configurations are exploited.

4.2.1.2 Economy, transparency and flexibility in compositional descriptions

To begin with, a compositional approach to describe discourse relations is more economic than a relation-based one, because the inventory of meanings to describe the seman-

tics of a same set of discourse relations is smaller, as can be seen in Table 4.2.

Moreover, a compositional approach allows to consider discursive meaning under a structuralist point of view. As in any linguistic system, discourse meaning can be represented as a structure organized by analogy, where few (if any) holes can be found. This implies that all possible combinations of the distinct meanings should be possible and equally used in the language, or at least comparably used. For example, if we find that a certain meaning is only combined with a subset of all the available labels in the proposed descriptive system, we can suspect that this kind of meaning is not of the same kind as the rest. Applying this kind of analysis allows to detect inconsistencies in a proposed configuration of meanings.

In Figures 4.1 and 4.2 we can see the structural representations of different configurations of the meanings used in Table 4.2. Each configuration highlights different relations between meanings. When we focus on the meanings *cause*, *parallel*, *revision* and *context* (Figure 4.1), we can see that there are some holes in the system, because some combinations of features never occur. When we focus in *continuation – elaboration* and *symmetric – asymmetric* meanings, we can find some systematicity to these holes: that the meanings of *symmetric* and *elaboration* almost never co-occur. This indicates a clear lack of economy in the system, which should be avoided.

Another interesting property of a compositional approach is the fact that the semantics of discourse relations are is transparent than in relation-based approaches. In both cases one has to resort to an inventory of pre-defined meanings to describe semantics. The main difference lies in the fact that, in a compositional approach, meanings tend to be more basic, more primitive than in relation-based approaches. They can be intuitively understood, and they are less prone to controversy than relations.

Furthermore, the fact that the semantics of discourse relations is partitioned in different meanings allows partial, underspecified descriptions, which are more feasible for shallow NLP techniques. We are discussing this in what follows.

4.2.1.3 Partitioning meaning to partition ambiguity

We can consider that a compositional approach to describe discourse relations partitions the meaning conveyed by these relations. Partitioning meaning implies partitioning ambiguity, so that the ambiguity of a certain discursive clue can be reduced to only part of its meaning, while the rest of the meaning can be asserted with a reliable degree of certainty.

We distinguish two kinds of ambiguity in the semantics of discourse relations: that which results of the incapacity of (shallow) NLP techniques to disambiguate the discourse semantics of textual clues (as in example (7)), and the inherent ambiguity of some discourse relations, which can very often present a vage, underspecified meaning (as in example (8)).

In example (7), the discourse marker *perquè* in Catalan is ambiguous between *reason* (*because, in case*) and *purpose* readings (*so that*). These two readings can sometimes be distinguished by the mood of the verb dominated by *perquè*: in *purpose* readings, the dominated verb must be subjunctive, while in *reason* readings it tends to be indicative.

cause			
cause	prog.		`result`
cause	prog.	sym.	*consequence*
cause	prog.	asym.	*purpose*
cause	elab.		`--`
cause	elab.	sym.	–
cause	elab.	asym.	*reason*
parallel			
parallel	prog.		`--`
parallel	prog.	sym.	*sequence*
parallel	prog.	asym.	*summary*
parallel	elab.		`--`
parallel	elab.	sym.	*restatement*
parallel	elab.	asym.	*example*
revision			
revision	prog.		`--`
revision	prog.	sym.	*contrast*
revision	prog.	asym.	*counterarg.*
revision	elab.		`--`
revision	elab.	sym.	–
revision	elab.	asym.	*concession*
context			
context	prog.		`--`
context	prog.	sym.	*narration*
context	prog.	asym.	*conclusion*
context	elab.		`background`
context	elab.	sym.	–
context	elab.	asym.	*circumstance*

cause			
cause	sym.		`--`
cause	sym.	prog.	*consequence*
cause	sym.	elab.	–
cause	asym.		`--`
cause	asym.	prog.	*purpose*
cause	asym.	elab.	*reason*
parallel			
parallel	sym.		`--`
parallel	sym.	prog.	*sequence*
parallel	sym.	elab.	*restatement*
parallel	asym.		`--`
parallel	asym.	prog.	*summary*
parallel	asym.	elab.	*example*
revision			
revision	sym.		`--`
revision	sym.	prog.	*contrast*
revision	sym.	elab.	–
revision	asym.		`--`
revision	asym.	prog.	*counterarg.*
revision	asym.	elab.	*concession*
context			
context	sym.		`--`
context	sym.	prog.	*narration*
context	sym.	elab.	–
context	asym.		`--`
context	asym.	prog.	*conclusion*
context	asym.	elab.	*circumstance*

continuation			
prog.	cause		`result`
prog.	cause	sym.	*consequence*
prog.	cause	asym.	*purpose*
prog.	parallel		`--`
prog.	parallel	sym.	*sequence*
prog.	parallel	asym.	*summary*
prog.	revision		`--`
prog.	revision	sym.	*contrast*
prog.	revision	asym.	*counterarg.*
prog.	context		`--`
prog.	context	sym.	*narration*
prog.	context	asym.	*conclusion*
elaboration			
elab.	cause		`--`
elab.	cause	sym.	–
elab.	cause	asym.	*reason*
elab.	parallel		`--`
elab.	parallel	sym.	*restatement*
elab.	parallel	asym.	*example*
elab.	revision		`--`
elab.	revision	sym.	–
elab.	revision	asym.	*concession*
elab.	context		`background`
elab.	context	sym.	–
elab.	context	asym.	*circumstance*

Figure 4.1:

symmetric			
sym.	prog.		`continuation`
sym.	prog.	cause	*consequence*
sym.	prog.	parallel	*sequence*
sym.	prog.	revision	*contrast*
sym.	prog.	context	*narration*
sym.	elab.		`--`
sym.	elab.	cause	–
sym.	elab.	parallel	*restatement*
sym.	elab.	revision	–
sym.	elab.	context	–
asymmetric			
asym.	prog.		`--`
asym.	prog.	cause	*purpose*
asym.	prog.	parallel	*summary*
asym.	prog.	revision	*counterarg.*
asym.	prog.	context	*conclusion*
asym.	elab.		`elaboration`
asym.	elab.	cause	*reason*
asym.	elab.	parallel	*example*
asym.	elab.	revision	*concession*
asym.	elab.	context	*circumstance*

continuation			
prog.	sym.		`continuation`
prog.	sym.	cause	*consequence*
prog.	sym.	parallel	*sequence*
prog.	sym.	revision	*contrast*
prog.	sym.	context	*narration*
prog.	asym.		`--`
prog.	asym.	cause	*purpose*
prog.	asym.	parallel	*summary*
prog.	asym.	revision	*counterarg.*
prog.	asym.	context	*conclusion*
elaboration			
elab.	sym.		`--`
elab.	sym.	cause	–
elab.	sym.	parallel	*restatement*
elab.	sym.	revision	–
elab.	sym.	context	–
elab.	asym.		`elaboration`
elab.	asym.	cause	*reason*
elab.	asym.	parallel	*example*
elab.	asym.	revision	*concession*
elab.	asym.	context	*circumstance*

Figure 4.2:

However, the semantics of the verbs in the main and subordinated clauses can override this constraint, as in (7-b), where the verb *declarin* is in subjunctive but the reading is not a purpose reading. The same pattern of moods, tenses and aspects of (7-b) can be seen in (7-c), but this example has a *purpose* reading, which we can infer from the relation between the predicates *demonstrate* and *declare impune*.

(7) Avui sento por perquè han declarat impunes tots els caps d'Estat.

a. Avui sento por perquè han declarat impunes tots els caps d'Estat.
Today I feel frightened because all heads of State have been declared impune.
b. Avui sento por perquè declarin impunes tots els caps d'Estat.
Today I feel frightened in case all heads of State are declared impune.
c. Avui em manifesto perquè declarin impunes tots els caps d'Estat.
Today I demonstrate so that all heads of State are declared impune.

Fully disambiguating the discursive meaning of *perquè* is beyond the current capabilities of NLP for Catalan. Within a relation-based approach, the only way to describe the semantics of this clue would be to associate it to a set of pre-defined labels; in this case, the relevant candidates would be *reason* and *purpose*. If we want to achieve a reliable degree of certainty in the resulting analysis, we will not be able to associate any of these labels to the clue, and we will not obtain any information from it. In contrast, if we are able to distinguish different meanings conveyed by this discourse marker, we may be able to determine some of the meanings with a reliable degree of certainty, and leave only some of them underspecified. If we follow the compositional meanings described in Table 4.2, we will be able to associate *perquè* to a *cause* and *asymmetric* meaning, while we will not determine its meaning in the *continuation – elaboration* dimension.

Besides, discourse relations present different densities of meaning. In many cases, their semantics cannot be described with all the richness allowed by the descriptive machinery, simply because they do not convey such rich meaning. For example, it is not clear whether the discourse relation between the two sentences in (8) is a causal, concessive, sequential or any other relation. A speaker could possibly interpret it as any of these relations, but it is also arguable that it can be interpreted as a very vague relation, which could be described with the meanings of *symmetric* and *continuation* of Table 4.2, with no specification of the *cause – parallel – revision – context* meaning.

(8) The weather was nice. I went to the gym.

Relation-based approaches can handle these cases of ambiguity in a very simple way: by adding to the set of relations the label that expresses the underspecified meaning (*asymmetric cause*, *contiguity*). We see two main problems in this approach: that meanings can be added to the inventory in an unprincipled way and that there is no explicit relation between closely related meanings. We will address methodological issues to determine a set of discursive meanings in the next section. Concerning related meanings, a compositional approach clearly elicits semantic relations, since related meanings share one or more of the

meanings that compose their semantics. The fact that the semantics of discourse relations is explicitly related allows to deal with ordered *types* of relations.

As follows from what has been exposed, the expression of underspecified meanings comes naturally in a compositional approach. This is very convenient for applied purposes, as it allows different granularities in the analysis, providing higher compatibility and portability of the resources based on such an approach to other frameworks.

In this section we have argued that describing discourse relations as a conglomerate of distinct discursive meanings is more economic and transparent than with the traditional atomic labels, and that it captures certain discursive relations that escape well-established, relation-based approaches.

We have discussed the need to distinguish different dimensions of discursive meaning, wherein submeanings can be distinguished. We have also explained how partitioning the semantics of discourse relations can contribute to increase the reliablity of the meaning obtained from shallow textual clues, because ambiguity can be reduced to only one part of their semantics. For these reasons, we believe that a compositional approach to the semantics of discourse relations is specially adequate for our framework.

Regardless of the approach, a basic questions remains: how many and which are the meanings that we will use to describe discourse relations? We want to determine the inventory of meanings systematically, avoiding *ad-hoc* solutions and extensive use of stipulation. Ideally, we would like to infer such inventory from empirical data, from psychological experiments (Sanders, Spooren and Noordman 1992), from introspection (Knott 1996) or from any other kind of linguistic evidence. We would like to systematize these meanings in an elegant, economic system, without any holes. And we would expect that this inventory provides satisfactory descriptions for all the range of discourse relations that we want to describe.

Unfortunately, this ideal system is far from what we can achieve in real systems. We will have to sacrifice descriptive adequacy to the capabilities of shallow NLP techniques, and then find a compromise between the structural elegance of the descriptive machinery and the reduced descriptive adequacy.

In the next section, we address the methodological question of how to identify, characterize and systematize these relations.

4.2.2 Determining an inventory of discursive meanings

Once we have decided that we will describe the semantics of discourse relations as a conglomerate of meanings, we have to address the question: which inventory of meanings do we use to describe relations?

The inventory of discursive meanings is a basic question for all theories of discourse relations. In many cases, this inventory is totally shaped by application needs, as in most natural language generation systems. In other cases, researchers have tried to come up with a methodology to induce a set of relations from linguistic or psychological data, in order to avoid *ad hoc* stipulations.

Hovy and Maier (1995) synthesize various inventories of meanings to describe discourse relations, and note that the adequate number of meanings is an unsolved question, with proposals ranging from two basic relations (Grosz and Sidner 1986) to more than 100 (Oates 2001; Carlson, Marcu and Okurowski 2003). For our purposes, a too coarse analysis would fall short to assess the targeted relevance and coherence relations between discourse entities or segments. However, a too detailed analysis would produce undesired ambiguity.

The question of how many meanings should be distinguished can be considered as a side-effect of the question of *which* and *on what grounds* discursive meanings should be distinguished. In most existing inventories, meanings are distinguished in an unprincipled way, which can make the inventory proliferate beyond what is reasonable and manageable, both automatically or by humans. Knott (1996) presents a sound criticism of such approaches:

> The extra claim in RST – that text is coherent by virtue of the relations between its intentions – is virtually unfalsifiable without a method for specifying what is to count as a relation in the first place.
>
> Even incoherent texts can be analysed according to the relations between the intentions in their spans. For instance, the text in (2.13) seems incoherent at first sight:
>
> (2.13) John broke his leg. I like plums.
>
> Yet we could still define a relation which holds between the intentions underlying the spans in this text: perhaps we could call the relation inform-accident-and-mention-fruit. [...] Clearly, we do not want to include these sorts of relations in any principled set of coherence relations.
>
> Knott (1996) (original numbering)

However, the general picture of inventories of discursive meanings is far from chaotic. As can be seen in the general picture provided by Hovy and Maier (1995), most inventories share a common core of basic coherence relations, and it is only in the detail that they differ. Concepts like *causality*, *negative polarity* or *elaboration* are present in almost any theory of discourse relations.

We believe that a certain degree of stipulation cannot be avoided in any theory about the organization of discourse, because any representation of discourse has a target, which biases decisions with respect to what kind, which or how many discourse relations will be distinguished. Given that bias cannot be avoided, it is convenient to make it as explicit as possible. It is also convenient to restrict bias to the minimum. In our case, the dimensions of discursive meaning that we were going to work with have been fully stipulated, but further divisions within these dimensions have been induced empirically.

Our main sources of bias are the representation needs of text summarization and the restrictions of shallow NLP. Since we are working with written text, we disregard the whole range of interpresonal and affective meanings that are mostly conveyed by discourse

particles in oral discourse. The fact that we do not make a deep analysis of the text implies that we do not model intentions, and that we only have shallow indicators of the thematic structure of the text. Finally, the most important source of bias is the fact that we focus on coherence and relevance, thus disregarding discourse relations that do not have any effect with respect to them.

Our *a priori* biases have determined the kind of discursive meaning that we are going to represent: they have determined that we are going to distinguish two dimensions of the meaning of discourse relations, intuitively corresponding to relevance and coherence assessment. The kind of meaning covered by each of these dimensions is detailed in the next section. In what follows we will describe how we have distinguished divisions of meanings within each dimension.

4.2.2.1 Establishing divisons of meaning within dimensions

We have tried to avoid subjectivity to establish divisions of meaning within dimensions. We have applied data-driven methods to induce relevant distinctions in the meaning of discourse relations, based on the evidence provided by discourse markers. In contrast with other researchers that have exploited this method, we have tried to constrain what qualifies as a textual cue to elicit a basic discursive meaning. We have relied on grammaticalization and cross-linguistic patterns of behaviour. This has led us to a coarse-grained inventory of meanings that nevertheless seems suited to meet our representation needs.

Ballard, Conrad and Longacre (1971) and Longacre (1983) claim that the existence of a discourse marker in a language serves as evidence of the existence of a particular type of interclausal relation. More cautiously, Martin (1992) says that discourse relations are markable by surface discourse markers. Knott (1996) determines a whole set of meanings to describe discourse relations based on evidence obtained from the inter-substitutability of discourse markers (or cue phrases, in his own terms). Many other authors make use of discourse markers to support or describe a set of discourse relations: they associate each of their proposed discourse relations to one or more discourse markers that prototypically signal it, and claim that a relation between discourse units holds whenever any of its associated discourse markers can be used to relate them (Hobbs 1985; Mann and Thompson 1988; Scott and de Souza 1990; Sanders, Spooren and Noordman 1992).

For example, following this method it can be argued that the *cause* relation exists because there is at least one discourse marker *because* that signals it, whereas there is no discourse marker to signal *inform-accident-and-mention-fruit*. Moreover, this signalling has to be distinctive: a certain relation (or feature of the meaning of relations) can only be said to exist when one can find a minimal pair of discourse markers that can only be distinguished by the proposed meaning.

Substitution tests have been successfully used (e.g. Knott 1996; Portolés 1998) to determine whether a given pair of discourse markers constitute a minimal pair: if they can be mutually substituted by each other in all contexts, they are in free distribution and do not constitute a minimal pair, but are signalling a single discursive meaning. If there is at least one context in which one of them cannot substitute the other, they present a

(partially) complementary distribution, which means that they are a minimal pair, and provide evidence to distinguish different meanings.

This method has a weak point: it crucially relies on the concept of discourse marker, whose definition is rather controversial. Knott and Dale (1996) proposes a test to determine when some lexical item is a discourse marker (described in Figure 3.3), but at least two criticisms can be made to it: that it only considers inter-clausal connectives and that it does not take into account grammaticalization or cross-linguistically recurrent polysemy. These two last aspects seem to be very valuable to identify **basic** discursive meanings. Indeed, there are many discourse markers and other kind of evidence that signal discursive relations, but we believe that only those that are highly grammaticalized signal basic meanings.

For example, Knott claims that a feature of the semantics of causal discourse relations is the fact that they rely on a semantic or pragmatic source of coherence. However, the discourse markers that make a distinction between these two kinds of causes (seen in example (9)) are much less grammaticalized than those that make the distinction between, say, negative or positive polarity relations (seen in example (10)).

(9) a. The footprints are deep and well-defined. { It follows that / So / # As a result, } the thief was a heavy man. *(pragmatic)*
b. I had a puncture on the M25 on my way back from work. { As a result, / So / # It follows that } I missed most of the first half. *(semantic)*

(10) a. Jim had just washed his car, { so / and / # but } he wasn't keen on lending it to us. *(positive)*
b. It was odd. Bob shouted very loudly, { but / and / # so } nobody heard him. *(negative)*

Under Knott's test, any set of words that relate two propositions seems to qualify as a discourse marker virtually. For example, expressions like *In considering this* or even *Taking carefully into account all this, which she had learnt during her long years at that dreadful school*, could also be considered discourse markers. Indeed, these expresions seem to signal relations between discourse units, as in the example below, but they do not seem to be indicative of **basic** discursive meanings.

(11) Further support for this concept has recently come from the work of Iftikhar who showed higher concentrations of bile in oesophageal aspirates from patients with Barrett's columnar lined lower oesophagus. In considering this, Stoker and Williams conclude that when gastric and duodenal secretions mix there may be a toxic synergism between the two that leads to mucosal disruption and intracellular damage to oesophageal cells.

In practice, Knott's test does not contribute to overcoming the methodological inadequacy of an open-bounded inventory of relations that has been criticized in purely stipulative approaches. If discourse markers are to be exploited to induce a set of basic discursive meanings, it seems necessary to distinguish those discourse markers that are signalling

precisely a basic meaning, among the set of discourse markers that signal any discourse relation. This does not imply that discourse markers that fall out of this more restrictive set of discourse markers do not perform a discursive function, but just that they constitute only weak evidence to support the existence of a basic discursive meaning. If this is not done, it becomes far too easy to find a set of words that serves as evidence to support a pre-determined relation, like *semantic* or *pragmatic* source of coherence. In addition, *ad hoc* exceptions have to be made for cases that pass the test but that he does not want to consider as discourse markers, like phrases that refer to the text where they are situated (*in the next section*) or phrases containing comparatives (*most surprisingly*).

4.2.2.2 Grammaticalization and cross-linguistic polysemy

We have tried to overcome the inconsistencies of Knott's test in two main ways. First, we have restricted the kind of meaning that we want to work with *a priori*, at the moment where the whole approach is conceived. As said before, we believe that a certain degree of stipulation is unavoidable in an application-oriented representation of text, but we try to make explicit and isolate the stipulative aspects of our approach. In this way, we avoid the *ad hoc* distinctions that Knott seems to make to sort out unwanted meanings, as seems to be the case with *in the next section*.

On the other hand, we have taken into account grammaticalization and cross-linguistic patterns of behaviour to restrict the set of discourse markers that qualify as sources of evidence to distinguish basic meanings, to sort out discourse markers like *most surprisingly*, but also like *it follows that*.

Grammaticalization is one of the main mechanisms to increase the economy of the linguistic system. Put simply, grammaticalization allows frequent meanings to be expressed as easily as possible, usually, in the shortest and most unmarked forms. Of course, in order to preserve the expressiveness of the language, this tendency to simplify linguistic forms has to be balanced with the distinctive power of these forms, that is, they still have to be able to convey as many distinct meanings as is necessary. Following the minimal effort rule, we expect that language provides highly grammaticalized forms to express the most frequent discursive meanings (which we will assume are the most basic), so that they can be expressed and understood with the minimum effort necessary. In contrast, since the size of the lexicon of a language has to be kept as small as possible, more sofisticated meanings are not expressed by items stored in the lexicon, but by linguistic forms that are created and understood compositionally, probably at processing time. As follows, the degree of grammaticalization of a discourse marker can be taken as an indicator of how basic a meaning it conveys.

Good indicators of grammaticalization are:

- the probability of occurrence of a lexical item,
- its length, and
- the degree of compositionality that can be observed in its meaning.

For example, we can say that the minimal pair *it follows that* and *as a result* in (10) are much less grammaticalized than the minimal pair *but* and *so* in (9) because the latter are much more frequent in text (108,339 and 35,871 occurrences of *but* and *so* in the BNC vs. 36 and 1,813 occurrences of *it follows that* and *as a result*), they are shorter, and because they do not have a transparent compositional meaning, which the first have.

Since most, if not all one-word discourse markers share the properties of *but* and *so*, degrees in grammaticalization can only be properly distinguished in multiword discourse markers.

The transparency of the compositional meaning of a multiword discourse marker can be approximated as the mutual information holding between its words. Then, the degree of grammaticalization G of a multiword discourse marker dm can be considered as directly proportional to the probability of occurrence of dm in text ($P(dm)$) and to the mutual information of the words that compose dm ($MI(dm)$), and inversely proportional to its length.

$$G(dm) = \frac{P(dm) * MI(dm)}{length(dm)} \tag{4.1}$$

For example, the discourse marker "*todo ello a pesar de que*" has a very low value of G because it is very long and the words that constitute it are very frequent and have very low mutual information. In contrast, the discourse marker "*sin embargo*" (*nevertheless*) has a much higher value of G because it is shorter and, although "*sin*" is a very frequent word, "*embargo*" tends to occur almost always after "*sin*", so they have a very high value of mutual information. Moreover, "*sin embargo*" is much more frequent in corpus than "*todo ello a pesar de que*".

Different granularities of discursive meanings can be distinguished by taking into account progressively less grammaticalized discourse markers. Highly grammaticalized discourse markers convey basic discursive meanings, and progressively less grammaticalized discourse markers make further distinctions within the basic meanings. This implies that, just like the methods discussed previously, our proposal is intrinsically open-bounded, and may yield an infinite inventory of discursive meanings. In contrast to other methods, however, the degree of grammaticalization allows to set an explicit bound to the inventory; a bound that can be quantified empirically.

It would be desirable, however, to determine the threshold degree of grammaticalization in an objective way. Cross-linguistic patterns of behaviour seem to provide the right kind of evidence to do that.

4.2.2.3 Semantic maps for the distinction of meanings

We have explained how *substitution tests* have been applied to identify distinct meanings based in minimal pairs of discourse markers (Knott 1996; Portolés 1998). These tests are also useful to establish ordered relations between discourse markers, reflecting the area of meaning they cover. These ordered relations are graphically represented in a (possibly

multi-parent) tree-like structure. In such a structure, a discourse marker covers all the meaning covered by the discourse markers under its yield.

In such a structure, basic meanings can be distinguished from non-basic meanings by taking into account only the distinctions introduced by discourse markers over a given threshold of grammaticalization. However, this representation does not provide a method to determine this threshold empirically. In contrast, semantic maps allow to determine this threshold by exploiting recurrent cross-linguistic polysemy (or multifunctionality, following Haspelmath 2003).

Semantic maps (Croft 2001; Haspelmath 2003) have been used to deal with data which cannot be properly described with existing terminology, to prevent terminological multiplication and to represent complex interactions of linguistic data more adequately.

Just like in one-language substitution tests, semantic maps exploit minimal pairs to determine an inventory of distinct meanings. In its simplest form, a semantic map consists of a number of grammatical functions plus a means to link these functions together, as appropriate. More than one language is typically taken into account to establish such inventory, so that substitution tests are applied cross-linguistically.

After this cross-linguistic inventory is established, it is organized in a two dimensional map, where meanings that can be conveyed by a single polysemous item (morpheme, word, collocation, in any of the languages considered) are adjacent, and all meanings conveyed by a given item (also in any of the languages considered) must occupy a continious area, wherein there is no meaning that cannot be conveyed by that item. The basic assumption in semantic maps is that cross-linguistically recurrent polysemy indicates relatedness in meaning, and that this relatedness in meaning can be represented topographically.

Semantic maps have been applied to various kinds of lexical items. The closest work to the application of semantic maps to discourse markers is that of Kortmann (1997) and Malchukov (2004), who have successfully applied the theory of semantic maps to formalize the cross-linguistic relations between the different meanings conveyed by subordinating and coordinating conjunctions, respectively.

An interesting property of semantic maps is that, as a result of cross-linguistic comparison, we can clearly distinguish areas of meaning that are common to all the languages considered. These areas can be taken as basic meanings. Haspelmath (2003) claims that, if enough not genetically related languages are used to build a semantic map, the areas of meaning that will emerge can be considered as universal meanings.

In our case, we are only taking into consideration three languages, and moreover, languages that belong to the Indoeuropean family: Catalan, Spanish and English. Therefore, we cannot claim that the areas that we can obtain by comparing the distribution of meanings in these three languages reflects universal patterns of meaning. However, the cross-linguistic distribution of the meanings of discourse markers in these three languages can help us determine the threshold in the degree of grammaticalization to distinguish discourse markers that signal basic discursive meanings from those that make finer-grained distinctions.

We consider that discourse markers that signal basic discursive meanings are those that fully cover a distinct area of meaning, with no overlap with any other area. In an area

of meaning with no overlap, all discourse markers with some meaning therein do not have any meaning in any other area. A discourse marker that fully covers an area of meaning can substitute any of the other discourse markers in that area, although the reverse may not necessarily hold.

We cannot avoid making an *ad hoc* exception here to exclude lexical items that may have a discursive function, but whose meaning is extremely vague, like *and*. We believe that they convey virtually no discursive meaning of their own, with a meaning closer to punctuation or mere adjacency than to content-rich discourse markers like *because*. Their vagueness allows them to act as place-holders for any relation, which could lead one to think that they can convey a very wide range of meanings. However, the meaning of these relations is not provided by the lexical item relating them, but by other mechanisms, like regular implicature.

The grammaticalization degree of content-rich discourse markers covering a non-overlapping area of meaning can be taken as the grammaticalization threshold to determine which discourse markers convey a basic discursive meaning. This threshoold is useful to identify discourse markers conveying basic meanings when areas are not clearly distinct. Since we are using only three languages to obtain semantic maps, most of the areas of meaning that we can obtain are doubtul, and the grammaticalization threshold is very useful to enhance the set of discourse markers that we can take as basic.

4.2.2.4 Organization of meaning within dimensions

Robust NLP approaches rely on default representations of the data. This default representation is provided whenever no other representation can be obtained. Any other analysis of the data is provided whenever we find evidence to override the default. Formally, this procedure can be considered an implementation of the defeasible inference approach that has been extensively used to obtain representations of discourse.

We have established a default representation for each of the two dimensions that we have distinguished. These default meanings are characterized as the absence of any of the other possible meanings in the dimension. They convey the unmarked implicatures for the continuation of discourse, and they are typically unmarked.

The range of possible meanings in each dimension is ordered in a hierarchy of markedness, so that the least marked meaning is the default, and the rest are only identified in the presence of adequate evidence. This hierarchy of markedness has been translated into decision trees, as those displayed in Figures 4.3 and 4.4.

In this section we have presented some criticism to common approaches to determine an inventory of discursive meanings to describe discourse relations. We have discussed how data-driven and application-driven approaches can be combined to motivate a useful inventory, that is both descriptively adequate and adapted to a certain framework. In the following section we apply this methodology to determine a set of discursive meanings to describe discourse relations.

4.2.3 A minimal description of discourse relations

In this section we apply the methodology discussed in the previous two sections to determine an inventory of discursive meanings:

- to describe discourse relations compositionally, as a conglomerate of basic discursive meanings,
- basic discursive meanings are organized in dimensions that distinguish heterogeneous kinds of discursive meaning and group together homogeneous kinds of meaning, restricting the possible combinatory of discursive meanings,
- dimensions of meanings are motivated by our representation purposes (relevance and coherence assessment),
- distinctions in meanings within dimensions are motivated by the empirical evidence provided by discourse markers,
- the whole system of meaning is structurally economic, none of the possible combinations of meaning is significantly underrepresented.

Since our purpose is to identify relevance and coherence relations in text, we have chosen to distinguish two main dimensions of the meaning of discourse relations: one that accounts mainly for relevance and another for coherence.

Cohesive relations (Halliday and Hasan 1976) have been taken as an indicator for the relevance of discourse units. If discourse is represented as a graph, where discourse units are nodes, those units that establish more relations with other units is more relevant. These relations have been typically identified by cohesive properties: discourse units that deal with a same topic are related (Salton, Singhal, Mitra and Buckley 1997; Morris and Hirst 1991; Barzilay 1997). In contrast, *coherence* relations serve to indicate the quality of the relationships between discourse units, be them cohesive or not.

However, as we showed in Alonso and Fuentes (2002) and Alonso and Fuentes (2003), these two kinds of relations provide an improved account of coherence and relevance relations if they are combined. We showed that an automatic summarizer that combined these two kinds of information outperformed a summarizer that exploited only one of them. Therefore, the two dimensions of meaning that we propose to distinguished are not labelled as *relevance* and *coherence*, but as *structural*, to account for cohesion relations, and *semantic* to account for coherence relations. We do not use the terms cohesion and coherence because the terms structural and semantic correspond better to the kind of information we will be actually exploiting.

The structural dimension describes the relations between discourse units as instances of the relations between units in a well-known type of data structure. We can consider that the semantic dimension is represented as labels of the relations determined by the structural dimension.

In what follows we will describe the kind of meaning that is to be covered by each of the two dimensions of discursive meaning that we have proposed to distinguish. Then, we will proceed to describe the subkinds of meaning that we have distinguished within each of these dimensions, and how they can be determined from a set of prototypical discourse markers in Spanish, Catalan and English (described in Section 3.4). We finish by describing the properties of the default meanings in each dimension.

4.2.3.1 Dimensions in the semantics of discourse relations

4.2.3.1.1 Structural dimension The structural dimension describes the relations between discourse units as instances of the relations between units in a well-known type of data structure. For example, if discourse is modelled as a stack (Grosz and Sidner 1986), the structural dimension describes whether a discourse relation makes the unit(s) under its scope to be pushed or popped from the stack. If discourse is modelled as a tree (Webber 1988; Polanyi 1988), discourse units are considered as nodes, and the structural dimension describes the relation between a node and the rest of nodes in the tree, with a special attention to the node where it is attached.

Under different forms, this kind of meaning has been central to many descriptions of discourse relations. Most of the aspects of the semantics of discourse relations that resort to the notions of topic or intention can transalted as instructions to insert a discourse unit in a data structure. For example, forward-looking speech acts (Cooper *et al.* 1999) can be translated to an instruction of pushing a new unit in a stack.

As can be expected, the variability of meaning within the structural dimension is highly restricted by the kind of structure whereby discourse structure is modelled. As explained in Section 3.2, we consider that discourse can be modelled as a sequence of local structures that can correspond to topics or intentions. Each topic (or intention) can in turn be modelled as sequence of subtopics (or subintentions), constituted by smaller topical (or intentional) units, until the level of minimal discourse units is reached. The RST claim that the rhetorical structure of a text can be modelled as a hierarchical tree that covers the whole text (Mann and Thompson 1988), only applies for texts that deal with a single topic (or intention).

In our implementation with shallow NLP techniques, we do not take the bottom-up procedure of building the structure from the leaves to its root, but the reverse: we start from minimal discourse units to build progressively bigger topical or intentional units. As explained before, the kind of evidence that we are working with allows to identify discourse structures at a low level of discourse, therefore we will model this dimension as a sequence of local structures, each structure covering the biggest topical or intentional unit that we can reliably identify, and describing the organization of any smaller units it may contain as a hierarchical tree. These smaller discourse trees are comparable to Polanyi *et al.* (2004)'s basic discourse units (BDUs).

In the structural dimension we target the representation of the topographical configuration of discourse. In order to obtain this configuration, we have to determine, for each discourse unit:

- its location in the structure of discourse, by determining the node in the structure (that is, the discourse unit) where it is attached
- the topographical relation with the node where it is attached, that is, the shape of the arc linking the two nodes (discourse units)

The first aspect does not specifically concern the semantics of the relation, but could be considered as its *syntactic* behaviour. We consider that the structural semantics of the relation between two discourse nodes is the shape of the edge that relates them. In a structure like the one we propose, this edge can have two shapes: either two units are at the same level, or they are at different levels. We present the heuristics to determine these two aspects of structural meaning in Appendix B, in this section we deal with the linguistic aspects of their semantics.

The semantics of an edge linking two discourse units at the same level is that the second contributes a **continuation** to the development of the discourse in an equal footing as the first one. In this respect, it is comparable to Grosz and Sidner (1986)'s *satisfaction-precedence*, RST's *multinuclear* or SDRT's *coordinating* relations.

The semantics of an edge linking two discourse units at different levels is that the one at the lower level **elaborates** on the information provided by the one at the higher level. It is comparable to Grosz and Sidner (1986)'s *dominance* relation, RST's *nucleus-satellite* or SDRT's *subordinating*. Typical examples of elaborations are examples, reasons, backgrounds, concessions, and a long etc. Just like in sentential syntax, discursive hypotaxis is more marked than parataxis, and therefore the inventory of hypotactic relations that a speaker has in mind can be much finer-grained than for paratactic relations.

Since structural relations configure the structure whereby we represent discourse, they also configure the *right frontier* of the discourse tree (Polanyi 1988; Webber 1988). The right frontier is the rightmost edge of a discursive tree, and it contains the nodes that are available for attachment of subsequent discourse units. Elaborative relations increase the number of nodes that are available for attachment of subsequent discourse units, that is, they enhance the right frontier, while continuative relations never add new nodes and they may even reduce the number of nodes that are available for attachment, if the attachment point is in a high level of the structure. These effects apply to a restricted graph representation as the one we propose, as well. These effects are displayed graphically in Appendix B, in Figures B.3 and B.4.

A good reason why structural meaning can be considered as distinct to other kinds of discursive meaning is the fact that they are often used as an orthogonal distinction in many proposals. Grimes (1975) or Martin (1992) use the paratactic/hypotactic distinction to classify discourse relations. Even in some implementations of RST, this distinction has been implicitly used as an orthogonal distinction. For example, Carlson, Marcu and Okurowski (2003) distinguish different kinds of causal relations that only differ in the fact that they are multinuclear or nucleus-satellite.

The distinction of meanings in the structural dimension is clearly influenced by the kind of representation of discourse that we use. However, discourse markers provide enough

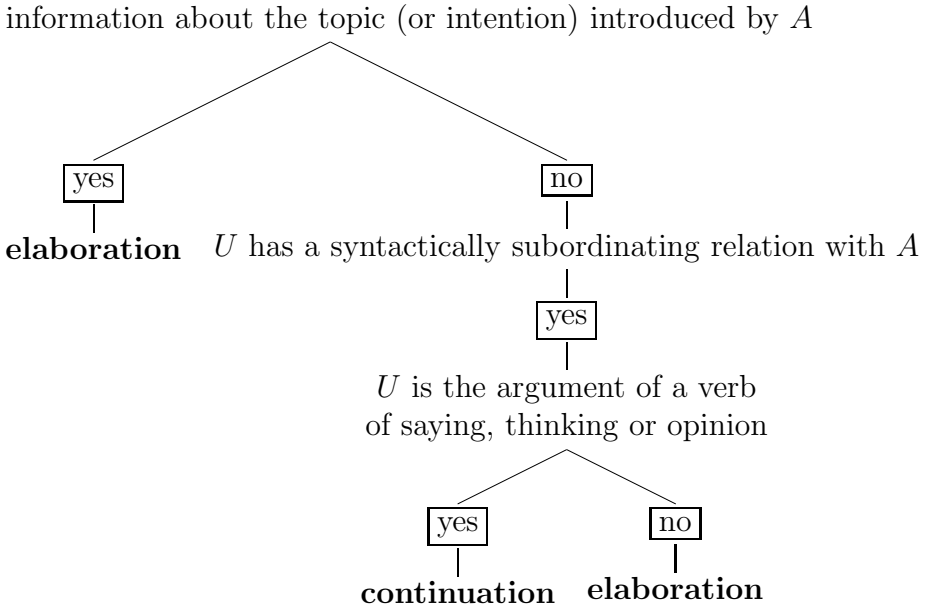


Figure 4.3: Decision tree to determine the structural meaning of the relation between a discourse unit U and the unit A to which it is attached.

empirical evidence to support this division of meaning. In Sections 4.2.3.2.1 and 4.2.3.2.2 we will discuss evidence that characterizes continuation and elaboration as distinct meanings.

As said above, robust NLP approaches rely on default representations of the data, that is provided whenever there is no evidence that signals a more informative relation. The range of possible meanings in a dimension is ordered in a hierarchy of markedness, so that the least marked meaning is the default.

As follows from the above discussion, the default meaning in the structural dimension is *continuation*, because is less marked than *elaboration*. *Elaboration* is typically signalled by sentential hypotaxis or co-reference, with one main exception: in reported speech, the syntactically subordinated element is discursively more relevant (Verhagen 2001), as in example (12), where the underlined segment holds an asymmetric relation with its matrix clause but it is discursively more relevant.

(12) [..] estoy convencido de que dentro de poco superará el examen de plancha, [...]
[...] I am sure that she will pass the ironing exam in short, [...]

In Figure 4.3 we describe the decisions that have to be taken to determine the structural semantics of the relation of a discourse unit with respect to the discourse unit to which it is attached. Note that it is easier to describe the relation of *elaboration* because it is more marked and therefore it is easier to identify what marks it, with the only exception of reported speech.

In sum, the structural dimension accounts for the relation of each discourse segment to the topics or intentions that are dealt with in the text, and represents this relations topographically, as a directed acyclic graph.

4.2.3.1.2 Semantic dimension The semantic dimension of discourse is typically represented as labels in the relations determined by the structural dimension, but this need not be the case. In some cases, a single discursive unit can be related by many more semantic than structural relations, as in example 3.1, where the node *d* establishes a semantic relation with nodes *b* and *c*.

Since they tend to co-occur, semantic relations are not usually perceived separated from structural relations, but they are qualitatively different. What is proper of subject-matter relations is that their contribution to the representation of discourse can be expressed propositionally, as *"A is the cause of B"*, *"A and B happened at the same time*, *"A is usually not expected from B"* and so on. In contrast, one cannot express continuation or elaboration propositionally.

Moreover, this kind of meanings can override the default inference processes that arise when two discourse units co-occur. In example (13-a), there is no semantic meaning in the relation between the two discourse units, so default inference applies, giving as a result something like "*we went home and then we had a party*". Any further interpretation (for example, causal) is subject to the beliefs and world knowledge of the person who interprets the utternace. However, if some semantic meaning is explicitly added to the relation, possible subjective meanings can be objectivized, as in (13-b), where the possible causal meaning is made objective. It is even possible to override default inference, as in (13-c), where the inference that the event in the first clause happened before the event in the second clause is overriden.

(13) a. We went home. We had a party.
b. We went home because we had a party.
c. We went home after we had a party.

Moreover, semantically rich relations seem to have an effect on the syntactical structure of sentences. Kehler (2002) shows that the coordinating structure constraint can be overriden by more general rules that can be explained by the presence of certain discourse relations, namely *cause* or *resemblance*.

As we have said before, we will determine the divisions in meaning empirically, but it is noteworthy that there is a core of relations that can be found in most theories of discourse coherence. The family of **causal** relations can be found in virtually all inventories of rich relations (cause, consequence, purpose, inference, reason, enablement, etc.), different kinds of **equality** relations are almost always to be found (parallelism, resemblance, etc.). Meanings involving **negative polarity** are also very common (concession, contrast, correction). The various approaches differ mainly in the distinctions they make within these core meanings, and the number and kind of other relations they may add to these.

Based on empirical evidence, we distinguish the following basic meanings in the semantic dimension:

where

DU is the discourse unit that is marked for a discourse relation *R*

A is the discourse unit where *DU* is attached by the relation *R*

revision signals that some content conveyed by *A* (propositionally, by regular inference, by implicature or by the default development of narratives) is denied by *DU* (e.g.: contrast, concession, correction).

cause signals that the occurrence of a fact conveyed in *DU* or the utterance of *DU* causes *A* to happen or to be uttered, or the reverse: that the occurrence or utterance of *A* causes *DU* to happen or to be uttered (e.g.: cause, consequence, purpose).

equality signals that it is relevant to establish some kind of equality between all or some of the propositional content of *DU* and *A*, or between them as units of language. The equality can be of the kind *A is a subkind of B*, *A is an example of B*, *A is the same as B* (e.g.: exemplification, rephrasing).

context signals that *DU* provides information about the circumstances where *A* happens or is uttered (e.g.: background, time, location, manner, condition).

Even if a discourse relation cannot be characterized by any of the above meanings, we assume that some kind of semantic relation can be established between *A* and *DU*, by the mere fact that they are contiguous in the structure of discourse. This relation vehiculates default inference, which may be sensitive to variations in domains or genres. Following Asher and Lascarides (2003), we call this kind of meaning **narration**, and we consider it the default meaning in the semantic dimension. However, it is qualitatively different from the rest of meanings in that its meaning cannot be expressed propositionally and cannot possibly affect the default inference processes between two discourse units, precisely because it constitutes the channel for this inference.

As in the structural dimension, we have formalized the markedness relations between the different meanings in the subject-matter dimension in the decision tree displayed in Figure 4.4. In this tree, we describe the decisions that have to be taken to determine the semantics of the relation of a discourse unit with respect to the discourse unit to which it is attached. Note that the default semantic relation, *narration*, can only be described by the absence of any of the features that characterize the rest of relations.

4.2.3.2 Meanings in the structural dimension

After describing the dimensions of discursive meaning, we will proceed to describe each of the meanings that we have distinguished within each of them. We will motivate these divisions in meaning by evidence provided by discourse markers in Spanish, Catalan and English. When relevant, we will make reference to discursive meanings that have been proposed in the literature.

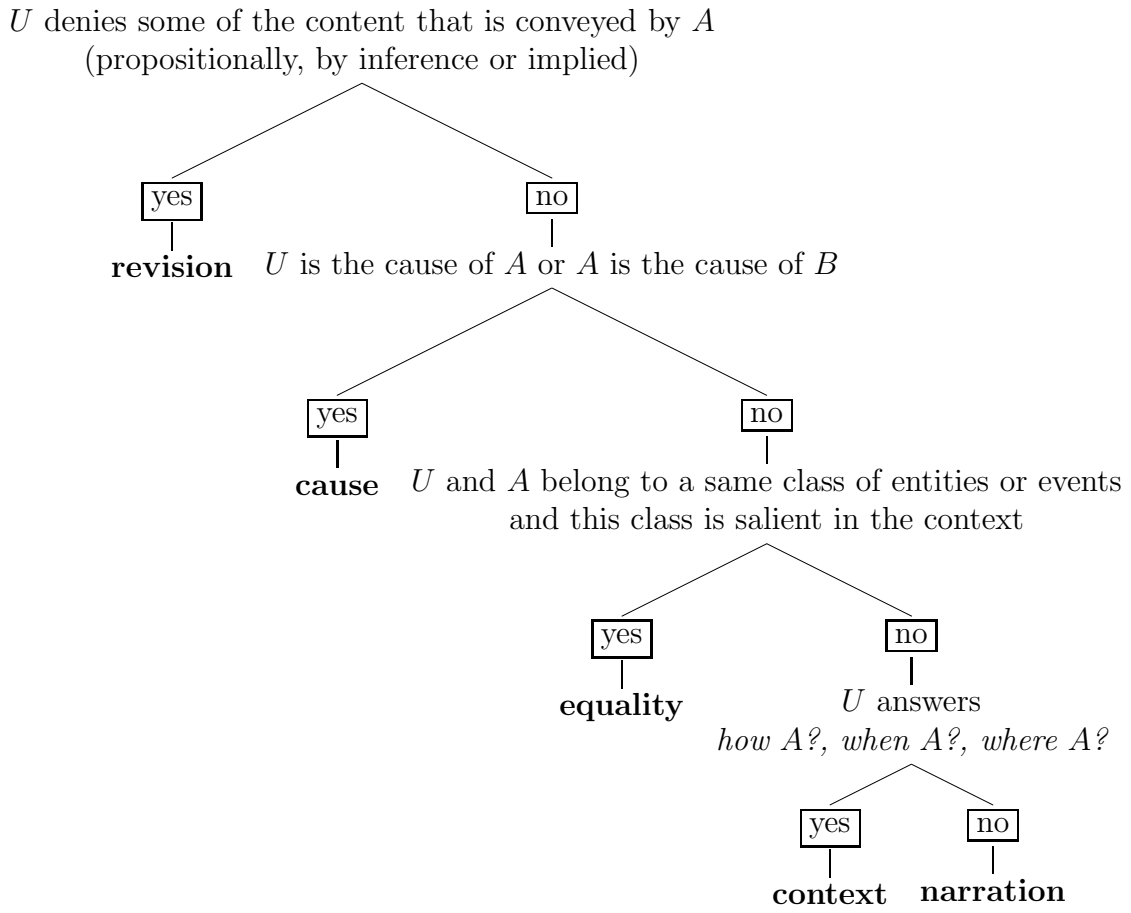


Figure 4.4: Decision tree to determine the semantic meaning of the relation between a discourse unit U and the unit A to which it is attached.

4.2.3.2.1 Continuation The semantics of an edge linking two discourse units at the same level is that the second contributes a **continuation** to the development of the discourse in an equal footing as the first one, as in Grosz and Sidner (1986)'s *satisfaction-precedence*, RST's *multinuclear* or SDRT's *coordinating* relations. Typical examples of continuation are lists, temporal or causal sequences of facts or events that are explained one after the other, as in the following example.

(14) In 1990 a police officer accused of distributing copies of a patriotic song to high-school students was sentenced to 13 years' imprisonment by a military court. Four men who allegedly planned demonstrations in December 1989 to commemorate the 1988 flag-raising, were sentenced to terms of between six and 12 years.

Continuation is the default meaning in the structural dimension, because what is by default expected of any discourse unit is that it continues the thread of the exposition. However, there are some discourse markers that mark precisely that something constitutes a continuation of the preceding discourse. In the examples below, the discourse markers do not convey any semantic meaning, but just express that the second discourse unit is a continuation of the first.

(15) a. At 13,000 pound plus options (like the CD), this is, short of the cabrio, the Escort/Orion flagship - thank heavens it has as much going for it as it does. The downside is a familiar one: notchy and sticky gearchange, mushy brake pedal, rough and vocal engine (this one had done 4000 miles yet would still not rev to the red line), wind noise at speed and,, the seats. When the new seats arrive, the Ghia and the S will be the first to get them.
So, it's okay by the standards of the rest of the range, but what about when faced with the real world?

b. We amateurs find various outlets for our products. At one end of the scale, some of use make things solely for our own use, or to give away to our friends and relations. Then there are those who sell their wares for charity at zero gain to themselves. As we progress up the scale of profit-taking, we find others who sell at small craft fairs and those who find an outlet through local shops, until we come to the favoured few who sell at the prestigious crafts fairs which Hugh prefers.

The discourse markers that can convey a purely continuative meaning are discourse markers that typically signal sequence, be it causal (*so*), temporal (*then*) or logical (*on the other hand*)[2]. However, parallel discourse markers in English, Catalan and Spanish show some discrepancies in their polysemy. The parallels of *so* and *then* in English in our discourse marker lexicon are *per tant* and *aleshores* in Catalan and *por tanto* and *entonces* in Spanish. However, the discourse markers *per tant* and *por tanto* are never underspecified to convey a

[2]Discourse markers with a presentational or logical meaning tend to behave like elements in a formal system, and are thus less prone to the ambiguity that is proper of natural language objects, and we will therefore leave them aside to study polysemy.

purely continuative meaning. It is the discourse marker *així* in Catalan and *así* in Spanish that can have this meaning. These discourse markers can convey consecutive and manner meanings, and purely continuative meanings, as in the following examples.

(16) El president nord-americà, Bill Clinton, es va mostrar complagut amb la contundent resolució expressada per l'ONU, perquè inclou alguns països que anteriorment havien mostrat "una lleugera tolerància" amb Saddam Hussein. Així, Rússia, que es va mantenir al marge de les condemnes internacionals en la crisi que l'Iraq va protagonitzar a principis d'any, s'ha mostrat ara en contra del desafiament.

(17) El Pla Especial de Protecció de Collserola que gestiona el Patronat (actualmente Consorci del Parc de Collserola) en el artículo 28 prevé la limitación progresiva de la caza mediante la declaración de figuras protectoras del territorio. Esta limitación se fundamenta tanto en la protección de la fauna como en la seguridad de los usuarios del parque. Por ello se consiguió en el año 1994 reducir las áreas de caza del 75% del territorio al 49%, incrementando las zonas de seguridad. Así Barcelona y Esplugues de Llobregat son dos municipios que fueron declarados, a petición propia, zona de seguridad en su integridad.

What seems to be crucial for discourse markers to convey a purely continuative meaning is their grammaticalization: *així* and *así* are shorter and much more frequent than the purely consecutive discourse markers, and this makes them more liable to have their meaning bleached away, so that they can be used to convey continuation only.

4.2.3.2.2 Elaboration The semantics of an edge linking two discourse units at different levels is that the one at the lower level **elaborates** on the information provided by the one at the higher level. It is comparable to Grosz and Sidner (1986)'s *dominance* relation, RST's *nucleus-satellite* or SDRT's *subordinating*. Typical examples of elaborations are examples, reasons, backgrounds, concessions, and a long etc. Just like in sentential syntax, discursive hypotaxis is more marked than parataxis, and therefore the inventory of hypotactic relations that a speaker has in mind can be much finer-grained than for paratactic relations.

(18) More than 130 political prisoners from Irian Jaya are currently serving lengthy prison terms for advocating the province's independence from Indonesia. Most have been convicted since 1988 under Indonesia's sweeping Anti-Subversion Law, accused of attempting to establish an independent state of "West Papua".

In contrast with continuation, it is hard to find discourse markers that convey elaboration only, without any added semantic meaning. Most discourse markers that convey purely elaborative meanings are polysemous between elaboration and continuation, like *moreover* in the example below.

(19) a. Can there be anything more telling about the deviousness of these people than his account of how they actually put on television and interviewed a man who

was said to have died while in a prison cell, and that, moreover, they did it with the sole motive of demonstrating that he was alive and in good health.

b. To scoffs of disbelief from some delegates, he asserted that, where managed safely and sensibly, nuclear power was one of the few energy sources which did not pollute the atmosphere. Moreover, to close all nuclear power stations would consign 100,000 workers to the dole queues.

Knott *et al.* (2001) argue that elaboration is not to be considered as the rest of RST-like coherence relations, because its semantics and linguistic realization are significantly different. They claim that it is possible to find at least one prototypical discourse marker by which all RST discourse relations can be realized in texts, but no such evidence can be found to systematically correlate with discourse relations conveying elaboration. Instead, elaboration seems to be realized via different kinds of discourse mechanisms, like reference, thematic trends in lexic, and also sentential subordination. This evidence on the linguistic realization of this relation seems to indicate that it does not have the same discursive effects as relations of *cause*, *concession* and the like, and, consequently, that it belongs to a different dimension of discursive meaning.

We feel that this discussion only makes sense wihtin a non-compositional approach to discourse relations. In our compositional approach, there are many discourse markers that signal some kind of elaborative relation, that is, that signal a discourse relation one of whose components of meaning is elaboration, as in the following examples.

(20) Time will tell whether there is widespread demand for the Discman and the other formats. In some respects they are clearly superior to normal books, for example they have database cross-referencing facilities ordinary volumes lack.

(21) 65 protesters were reported to have been burned to death when security forces set fire to a shopping centre in which they were seeking refuge.

There is a strong correlation between sentential subordination and discursive elaboration. Sentential subordination usually functions as discursive subordination as well, to the point that mechanisms of sentential subordination can extend their original uses to indicate discursive subordination, as is the case of Yup'ik, a language of the Eskimo-aleut family where discursive subordination is marked with the same devices as sentential subordination (Lasswell 1996). In the languages that we are working with (Catalan, Spanish and English), most discourse markers that convey an elaborative meaning are subordinating conjunctions. It is then productive to apply the defasible inference that any hypotactic syntactical relation corresponds to an elaborative discourse relation.

However, there are some notable exceptions to this general rule: reported speech (and all verbs of thinking or opinion), result-oriented causes (purpose, consequence) and rephrasing. This lead us to to distinguish a further kind of meaning in addition to structural and semantic, namely syntactic (Laura Alonso i Alemany *et al.* 2004). This meaning accounted for properties of relations that were straightforwardly derived from sentential syntax. This additional kind of meaning allowed us to provide an adequate de-

scription of the above cases: constituents introduced by a reported speech constructions were characterized as a *continuation* + *subordinating*, rephrasing was characterized as *elaboration* + *coordinating*. Different kinds of causes could also distinguised: *reason* was characterized as *elaboration* + *subordinating*, *purpose* as *continuation* + *subordinating* and *consequence* as *continuation* + *coordinating*.

As displayed in Tables 4.1 and 4.2, the structural of these meanings was highly antieconomic. Moreover, we found that the combination of those meanings seems descriptively adequate, but does not produce distinctions that are useful for relevance or coherence assessment. A discourse relation is equally continuative or elaborative irrespective of its syntactical realization.

The effects of sentential syntax on coherence can be directly obtained from the sentential analysis, without any discursive meaning involved. A subordinated construction (clauses or phrases) is usually related to its matrix clause, so that separating them makes the subordinated clause incoherent or even ungrammatical, and sometimes the matrix clause as well. But these effects do not affect the configuration of discursive structure, and their interaction with discursive meanings are very restricted, if any.

4.2.3.3 Meanings in the semantic dimension

4.2.3.3.1 Revision The contribution of the meaning of revision to the representation of discourse can be propositionally expressed as:

> *DU* denies some content conveyed by *A*, either propositionally, by inference or by implicature

This meaning has been widely studied in various fields, because there seem to be many cognitive and linguistic processes involved in the realization of this kind of meaning. From the point of view of cognitive processes, it is a costly procedure, from the point of view of argumentation, it is a complex move that implies a competition between arguments, usually where one of them wins, from the rhetorical point of view, it is a very effective strategy of communication, from the linguistic point of view, it is a highly marked structure in language, usually accompanied by side phenomena like marked modality, veritative expressions and irrealis.

Many different kinds of meanings have been distinguished within what we globally call *revision*: correction (22), concession (also called denial of expectation) (23) counterargument (24), semantic based contrast (25), or topic based contrast (26), Only in the case of correction is the negated content explicitly stated. In the rest of cases, the content that is negated is implied from the first argument of the revision relation by different mechanisms: in the case of concession, by default implicature, in the case of counterargument and semantic contrast, by structural opposition of the arguments, in their intentional or denotational content, respectively, and finally, in the case of topic based contrast, by the opposition between the default narrative action of continuing in the same line as the previous utterances and the marked action of introducing a change in the narrative.

(22) The Zen foot is not yellow, but white. The skin not waxy, but translucent; the texture not oily, but clean and dry.

(23) A state of emergency was declared but another 40 people were killed the following day.

(24) Not much flavour, but a fibrous-to-spongy chewiness which sets it the average savoury snack.

(25) Torture of political detainees in Mauritania has been routine since 1986, but it has never before been used on such a scale.

(26) To outsiders, the two events may seem little more than an international organization at work, Amnesty members these events are charged with significance. The Cold War was at its height when Peter Benenson, the British lawyer, founded Amnesty, and three decades later it is hard to believe that the Moscow AI Group finally has permission to become part of Soviet life. Similarly, the idea that a human rights concert should be held in the very stadium in Santiago where Allende's officers rounded up thousands of Chileans in 1973, prior committing gross violations, stretches the powers of credulity.
But incredulous, or not, the events happened (millions of television viewers worldwide watched the Chilean concert) and as such they typify the massive changes that Amnesty has undergone in its 30-year history.

As can be seen in the above examples, the discourse marker *but* can signal any of the meanings that have been distinguished in previous approaches. Its equivalent in Catalan (*però*) and Spanish (*pero*) can have exactly the same functions. This discourse marker is very frequent in text, and there is no shallow clue that characterizes any of these meanings.

Languages like Russian express different shades of revision meaning by different discourse markers (one signalling semantic opposition or pure contrast, and another signalling concession) that seem to be in complementary distribution[3]. In any case, we are only taking into account evidence from English, Catalan and Spanish, and in these languages there is clear motivation to distinguish a single kind of revision meaning.

This meaning is the most marked in the semantic dimension, it is usually heavily marked in text, by various means: evidentiality, negative polarity, degree markers and co-occurrence of various discourse markers with revision meaning, as can be seen in the following example:

(27) "Guess The Weight Of The Donkey" and "Pin The Tail On The Jam Sponge" are not, in fact, recognised PR mechanisms, contrary some reports, although it is true that they inspired the electoral systems in Israel and Italy respectively.

[3]Although this complementary distribution has to be taken with some reserve: Russian argumentative schemes may tend to avoid strong negations like the one conveyed by *no*, and rephrase sentences to use contrastive *a* instead

Precisely negative polarity is one of the features that has been usually associated with the kind of discourse relations that we have grouped under revision (Sanders *et al.* 1992; Knott 1996; Lagerwerf 1998). However, there are revision relations that involve no overt negative polarity, as in example (28). However, even in these cases, the meaning we have proposed for revision still holds, because the second clause denies some of the content conveyed by the first clause. In this example, the content that is denied is the regular implicature that, following the maxims of relevance and quantity, one always tries to be as informative as possible. This expectation is not fulfilled in the example, because the second clause is more informative[4] than the first one with respect to a relevant aspect of the first clause, namely, that there is someone that is fun.

(28) Maria is fun, but with Santi I can't stop laughing.

Moreover, there are some discourse markers that have inherent negative polarity, like *unless*, but do not convey a revision relation. Therefore, overt negative polarity does not seem to be necessary for revision relations.

4.2.3.3.2 Cause The contribution of the meaning of cause to the representation of discourse can be propositionally expressed as:

> the occurrence of a fact conveyed in *DU* or the utterance of *DU* causes *A* to happen or to be uttered, or the reverse: that the occurrence or utterance of *A* causes *DU* to happen or to be uttered

In contrast with revision, causal relations are not highly marked in text, usually there is only one discourse marker to mark them, but it is also very usual that no discourse marker is found, and that the relation is recognized by the speaker by inference. A good proof of the fact that causal relations are mainly establised by the content of the discourse units they relate are the works of Soricut and Marcu (2003) and Hutchinson (2004), who apply machine learning techniques to characterize rhetorical relations by the features of the discourse units they relate, obtaining remarkable results for causal relations.

Causal relations are distinguished in any approach to describing discourse relations that has more than two relations. Different kinds of cause have been distinguished, based on different features of their meaning. The most widely spread distinctions are:

- volitional vs. non-volitional (Mann and Thompson 1988)
- result-driven vs. cause-driven (also anchor-based vs. counterpart-based), depending on where the focus (or nucleus) of the relation is located (Mann and Thompson 1988; Knott 1996)
- pragmatic vs. semantic source of coherence, distinguishing whether discourse units are related because of their illocutionary force or by their propositional content, respectively (Sweetser 1990; Sanders, Spooren and Noordman 1992)

[4] If we consider grades of adjectives like quantifiers.

We find that the discourse markers that are most representative discourse for causal meaning are *so*, *because* and *(in order) to*, because they are highly grammaticalized. The kind of causal meaning conveyed by these discourse markers is very different, but relevant distinctions between them can be made by resorting to structural meaning: we can distinguish between *elaborative cause* (reason, *because*) and *continuative cause* (purpose, *in order to*, consequence *so*). This distinction is comparable to the result-driven vs. cause-driven distinction, but in our case we are not concerned about the location of the cause or the result of the relation, but rather about the location of the focus, as it is structurally signalled by discourse markers.

Even if different causal meanings can be distinguished and have been distinguished in the literature, discourse markers provide evidence that cause is to be treated as a compact core meaning. In Catalan, *perquè* and its phrasal counterpart *per* are ambiguous between a causal reading (*because*, *because of*) and a purpose reading (*so that*, *to*). This provides enough evidence not to distinguish reason from purpose. Then, discourse markers that typically signal consequence can also signal purpose, as in the following example.

(29) If part of the credit for a breakthrough could be claimed by Mr Hussein, and so induce him to pull his army Kuwait, why not solve two problems in a single bloodless stroke?

In sum, we believe that discourse markers provide enough evidence to treat causal meaning as a whole. As said before, this does not mean that further meanings cannot be distinguished within the causal domain, but we believe that this level of abstraction is useful to obtain a reliable representation ot causal relations via shallow NLP.

4.2.3.3.3 Equality The contribution of the meaning of equality to the representation of discourse can be propositionally expressed as:

> *DU* and *A* are equal as units of language, units of content or it is relevant to establish an equality between some part of their content.

The kinds of meaning that we consider under parallelism are all modalities of equality, of the form *A is mod(B)*, where the modalizing factor can be *is a kind of*, *is similar to*, *is the same as*, etc. This covers meanings like exemplification, comparison, logical structures (lists, two-sided arguments, etc), summary or rephrasing. We exemplify these usages below.

(30) In some respects they are clearly superior to normal books, for example they have database cross-referencing facilities ordinary volumes lack.

(31) It's not much fun, you're chained to the wall - it's like going back to a medieval library.

(32) Now the women want equal pay. What is this? Will they never be satisfied? On the other hand they are telling us we are not playing for the money but the love of the sport.

(33) Nevertheless, the review represents substantial progress. It holds out the promise of swifter verdicts. It appends that promise to a clear and comprehensive code of conduct (one that builds helpfully on the recently published Fleet Street set of principles). And it adds to the sanctions the Council may take - positioning of judgements in offending papers; a privacy hot line; and, more controversially, hauling in proprietors to discipline editors.
The Council , in sum, is seeking most of the tools it needs to relaunch itself as an active, respected self-regulator.

(34) In the past, some purists have said that all surface decoration applied to designed objects - all ornament, in other words - must be based on motifs which look flat.

The semantic map of equality is a rather interesting one, because it has the form of an overlapping continuum: the discourse marker *as* (note that it is highly grammaticalized) covers the meanings of *exemplification* and *comparison*; *also* covers comparison meaning and can be used to organize logical structures; summary and rephrasing can be expressed by the same discourse markers, discourse markers like *in essence* that show the opposite meaning as *for example*, that is, that come right next to it in the continuum of meaning.

It is clear that this continuum can be segmented because the rest of semantic meanings are clearly delimited, and also because we expect that the general mechanism of analogy is reflected at discursive level as it is reflected in other levels of language.

4.2.3.3.4 Context The contribution of the meaning of context to the representation of discourse can be propositionally expressed as:

> *DU* provides information about the circumstances where *A* happens or is uttered

Context can be regarded as the textual realization of the figure–ground principle of cognitive grammar. Langacker (1987) claims that it is a general pattern of human perception that some figure is made relevant by contrast to a background, and that this pattern is reflected in language. Indeed, there are plenty of ways in language to provide information about the context for a given entity or event.

Maybe because it is such a common use of language, many different kinds of context can be distinguished: time (point, duration, etc.), place (location, coordinates), manner, general background, condition (also known as enablement), etc.

It is not costly to distinguish temporal from spatial uses of discourse markers by shallow NLP techniques. One can quite reliably detect whether the content of a discourse unit dominated by an ambiguous discourse marker is temporal or spatial, simply by collecting a bag of location words (e.g.: *country, house*) and another bag of time words (e.g.: *day, year, period*) and applying pattern-matching techniques. However, this kind of information is qualitatively different from the information that we are exploiting in this thesis, so we will leave this line of research unexplored.

For coherence and relevance assessment, we believe that all these kinds of meaning behave in a comparable way. Moreover, discourse markers seem to provide evidence that a general "context" meaning exists. Most temporal prepositions (that can also act as subordinating conjunctions) are polysemous between a temporal and spatial readings, as seen in the example below[5].

(35) a. The releases occurred shortly before two AI representatives arrived in Swaziland for talks with the government.
b. Har Govind, a former chief commissioner of income tax, the existing limit of 40% will be increased to 51% under industrial legislation put before the Indian Parliament last month.

Some recurrent patterns of polysemy seem to suggest that there is a common meaning covering both cause and context. Discourse markers that signal temporal sequence (*then*) very often can also signal causal sequence. Moreover, discourse markers of manner are the preferred markers for consequence (*so*, *in this way*).

From the comparison of Catalan, Spanish and English, we have found no strong evidence to support a distinction between these two areas. However, there are descriptive reasons to prefer that these two meanings are distinguished. Indeed, the discursive effects of context with respect to relevance and coherence are very different from those of cause, since cause tends to establish stronger coherence effects than context. Moreover, there are various discourse markers exclusively signalling context or cause whose grammaticalization index is equal to that of the discourse markers that very clearly cover a whole area of discursive meaning, like *but*.

A clear definition and delimitation of the area of meaning we have called "context" needs to be further pursued, with evidence from other languages and probably also diachronic data. In the current work, descriptive adequacy has been the only strong reason to shape this area of meaning.

4.3 Discussion

In this chapter we have proposed a treatment of the meaning of discourse relations that is specially suited for text summarization via shallow NLP techniques. Our target was to determine which meaning could be reliably obtained via shallow cues and how it could be better represented.

We have described some linguistic phenomena that provide information on the structure of utterances beyond clausal level, and which can be reliably analyzed by shallow NLP techniques. We have focused in those phenomena that are especially adequate to our NLP capabilities, namely, punctuation, shallow syntactic analysis and discourse markers.

[5]We are not considering here the information introduced by very vague prepositions, like "*at*" or "*by*", just as we did not consider that the polysemy of *and* qualifies to identify basic meanings, because the lexical item is too vague, and we can consider that it conveys no meaning of its own.

name of the relation	semantic meaning	structural meaning
contrast	revision	continuation
concession	revision	elaboration
result-driven cause (purpose / consequence)	cause	continuation
reason	cause	elaboration
parallelism	equality	continuation
exemplification	equality	elaboration
topic	context	continuation
background	context	elaboration
narration	narration	continuation
explanation	narration	elaboration

Table 4.3: Discourse relations that can be obtained from the combination of the meanigns that we are able to distinguish reliably via shallow NLP techniques.

In the first place, we have argued that the meaning of discourse relations is best described compositionally. Then, we have determined two different kinds of discursive meaning that we find useful to determine relevance and coherence relations in text, namely *structural* and *semantic*. We have organized these two kinds of meaning as dimensions where distinct meanings have been distinguished relying on the evidence provided by discourse markers. These distinct meanings have been organized in a hierarchy of markedness, formalized in decision trees.

In sum, we have provided a set of seven distinct meanings (two structural and five semantic), organized in two dimensions of meaning, which allow to distinguish the ten basic discourse relations shown in Table 4.3. As will be discussed in Section 5.3, these relations seem to correspond to some patterns underlying human judgements on discourse.

These meanings are organized in a hierarchy of markedness supported by agreement of human judges in identifying relations between discourse segments. This organization significantly reduces the cost of taking decisions for a discourse parser: in case there is contradictory evidence, the most marked wins. What needs to be investigated is the relation between meanings and other kinds of evidence, like syntactic structure, punctuation, lexic... because these certainly determine the strength of a given evidence to mark some meaning.

Future work includes enhancing the scope of the inventory presented here, testing its applicability for uses other than text summarization. A very interesting line of research consists in integrating evidence provided by discourse markers in other languages, in relation with the proposed inventory of discursive meanings. This kind of evidence can contribute to settle an inventory of discursive meanings of more general scope and also to establish finer-grained divisions within areas of meaning that cannot be properly segmented with evidence from Catalan, Spanish and English alone.

Chapter 5

Empirical Support

In this chapter we provide empirical support for some of the theoretical concepts proposed in the previous two chapters. We base this support in two kinds of experimental data: evidence from human judgements and evidence from automatic procedures based in shallow textual clues. We provide evidence to support the concepts of discourse segment, discourse marker and the discourse relations that we are working with.

In Section 5.2, we present an experiment where we asked some human judges to summarize newspaper articles by removing words that they considered irrelevant. We assessed the agreement between judges with respect to the relevance of words, and found that judges do not agree on which words to remove from a text more than could be expected by chance, which is not surprising taking into account previous work in evaluation of summarization (see Section 2.2), and seems to reflect the inherent subjectivity of the task as it is defined.

However, a discretization of text in discourse segments and the classification of discourse segments in diferent classes allows to model the behaviour of judges so that interesting patterns of agreement can be found, thus supporting our hypothesis that text is segmented and that segments can be classified into different types with different discursive behaviours. We have shown that the probability that a word is removed can be predicted with a higher reliability if it is conditioned to the form and semantics of segment where the word occurs. For example, words occurring in a segment of context marked by punctuation and discourse markers have a higher probability of being removed than words in a relative clause.

But what is more interesting is that the level of agreement between judges is significantly higher for some kinds of discourse segments distinguished by their semantics. Agreement was much higher than the average when it concerned the relevance of words belonging to segments with strong discursive meanings, like *cause, equality* or *revision*. This supports our proposal of discourse semantics, which seems adequate to capture speakers' intuitions about the relevance of words, both for the unit of segment and for the categorization of segments. It also seems to support the hierarchy of markedness in discursive meanings, because there is more agreement for cases with more marked meanings.

The behaviour of judges, modelled according to the proposed organization of discourse, has provided very valuable data to refine summarization heuristics based on discourse representations. We have applied this information to identify and characterize relations between discourse segments as an aid to produce automatic summaries of e-mail messages in Carpanta (Alonso *et al.* 2003b).

Once we know that the concept of discourse segment has an empirical basis, we assess the empirical basis of relations between segments. In Section 5.3 we present an experiment where three judges (two naive judges and a linguist) identify relations between discourse segments that have been previously identified, and describe the semantics of these relations. The level of agreement between these judges does not allow to assert the stability and reproducibility of the task, but are well beyond chance agreement and support our claim that these features can capture the intuitions of speakers with respect to discourse organization.

Finally, we show that our theoretical proposal can be satisfactorily treated by a shallow NLP approach. In Section 5.4 we show that two algorithms to identify discourse segments obtain good levels of performance exploiting different kinds of shallow cues. This experiment also shows that discourse markers significantly improve the performance of the algorithms. Then, in Section 5.5 we describe a tool that relies on shallow linguistic evidence and large amounts of corpora to automatically enhance a starting lexicon of discourse markers.

First of all, in the next section we discuss how the significance of empirical data has been assessed: by studies of variance, hypothesis testing, measures of agreement between judges and measures of performance of automatic procedures.

5.1 Assessing the significance of empirical data

If empirical data are to provide support to theoretical claims, it is important to assess the significance of the data with respect to the hypotheses to be tested. Imagine we want to test whether two kinds of segments have different behaviours with respect to relevance, for example, segments within causal relations and within equality relations. To assess the different behaviour with respect to relevance, we have a corpus where each segment has been associated to a relevance score. Then, before finding differences or similarities in the scores associated to causal or equality segments, it is necessary to take into account the reliability of the association between segments and a score of relevance. One way to do that is to compare the actual association with a random association, and check whether they differ significantly, by **hypothesis testing**. Another possibility is to have more than one association for the same set of segments, produced independently, for example, by different people or different algorithms, and check whether their **ratio of agreement** is significative, typically, beyond chance agreement. We are going to use both these approaches (hypothesis testing and ratio of agreement) to assess the reliability of the data whereby we support our theoretical claims, and also the stability and reproducibility of the results presented here.

In our experiments, the data consist of a classification of linguistic units into predefined classes that we find relevant for summarization. For example, in Section 5.2 words are classified as removable or non-removable according to their relevance in a text. In Section 5.3 discourse segments are classified by their semantic features, assigned from a closed set of possible features. In Section 5.4 words are classified as belonging to the same segment as the word preceding them in the sequence of text or to a different segment. In

Section 5.5 sequences of words are classified as making part of a discourse markers.

We are not going to provide here an assessment of the significance of the results produced by automatic procedures, because the algorithms that produce the data are transparent and the hypothesis that they have been produced randomly can be rejected with a high degree of certainty. But assessing the significance of human judgements is more interesting, because there are more than one judge interpreting the same phenomena, and because the procedure whereby judgements have been produced is hidden. So, it needs to be tested whether there is a significant relation between the interpretation of different judges. In what follows we are going to discuss the methods we have exploited for significance testing as applied to human judgements.

5.1.1 Study of variance

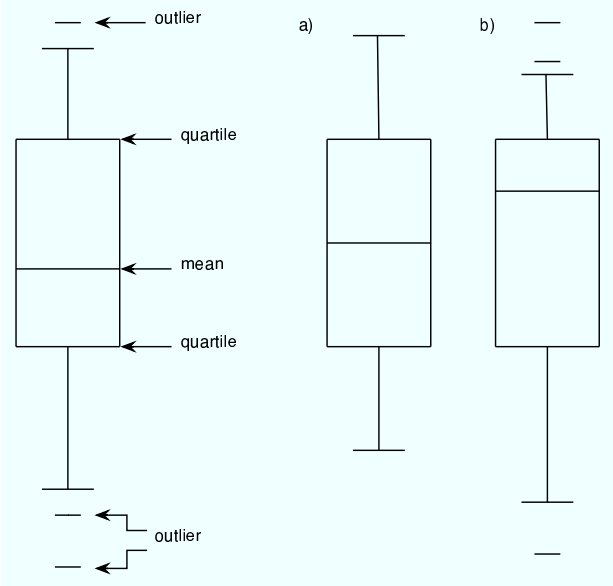


Figure 5.1: An example of how the variance of a given population with respect to a numeric parameter can be represented graphically.

In order to gain a first impression of the variability in judgements, that is, to know whether judgements by different subjects were comparable, we studied the distribution of values for relevant aspects of the judgements. We carried a study of variance of aspects like: amount of words removed from texts by each judge, amount of words removed in each text in average, amount of segments classified in each category, length of segments, etc. This study of variance consisted in obtaining the mean, the quartiles and identifying outliers, and we represented it graphically by a boxplot, which is interepretd as seen in the example in Figure 5.1. In this example, *a)* has been produced by a normal distribution, and shows less variance than *b)*, whose distribution is clearly not normal and shows much more variance than *a)*, since the tails of the boxplot are longer and it even has some outliers.

Boxplots also allow to get a grasp of the distribution of the data, which is a crucial aspect to choose tests of hypotheses. Most tests of hypotheses assume a certain distribution of the data in order to formulate hypotheses against which to test the distribution of the actual data, so it is necessary to know wether the actual distribution of the data matches the assumptions made in the test. Some measures of agreement, like the percentage of agreement between judges but also some computations of the Kappa coefficient, are also affected from prevalence an bias in the data, so it is important to inspect the data previously and see if the distribution of the values will produce any of these effects.

5.1.2 Hypothesis testing

In the first place, we compared the distribution of human judgements on the same set of items against a random distribution, to test whether the agreement of subjects on the words to be removed could be attributed to some underlying organization of text (alternative hypothesis) or it did not significantly differ from what can be obtained by chance (null hypothesis). We applied the t-test of difference in means to determine whether the difference between human and random judgements was significant enough or not.

The standard t-test (5.1) looks at the mean and variance of a sample of measurements, where the null hypothesis is that the sample is drawn from a *normal* distribution with mean μ. The test looks at the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely one is to get a sample of that mean and variance assuming that the sample is drawn from a normal distribution with mean μ.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{5.1}$$

where

N is the sample size
$\bar{x}$ is the mean of the sample
s^2 is the variance of the samples $var = \frac{\sum (x-\bar{x})^2}{N}$

In our case, we cannot assume normality of the samples, or at least not from the random sample, which is typically uniform. Therefore, we have chosen a variant of t-test that does not require the samples to have a normal distribution: the *matched samples* t-test, calculated with the variant of the t-test for correlated data (5.1) provided in the R package for statistical computing (The R Project for Statistical Computing 2004), which relies on the assumption that the data are correlated rather than assuming that the two samples are independent normal samples.

$$t = (\bar{X} - \bar{Y}) \sqrt{\frac{n(n-1)}{\sum_{i=1}^{n} (\hat{X}_i - \hat{Y}_i)^2}} \tag{5.2}$$

where

X, Y are the samples to be compared
$\bar{X}, \bar{Y}$ are the means of the samples
$\hat{X}_i = (X_i - \bar{X})$
$\hat{Y}_i = (Y_i - \bar{Y})$

Since the direction of the comparison of real and random distributions is irrelevant, the critical value of t is determined by a two-tailed distribution. Since the distribution of t is symmetric, the critical value of t is the same for one- and two-tailed tests. For infinite degrees of freedom and $p = 0.5$, the critical value of t is $t_{crit} = \pm 1.645$[1], meaning that values of t bigger than 1.645 and smaller than -1.645 indicate that the distributions differ significantly and so the null hypothesis that they are the same distribution can be rejected, allowing that this is not the case for 5% of the sample. A detailed relation of critical values of t for various values of p can be seen in Table 5.1.

degrees of freedom	0.10	0.05	0.025	0.01	0.005	0.001
100	1.29	1.66	1.98	2.36	2.63	3.17
∞	1.28	1.64	1.96	2.33	2.58	3.09

Table 5.1: Critical values of t for varying degrees of freedom.

5.1.3 Ratio of agreement

We studied correlation by applying Intraclass Correlation (ICC), which assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects, with the following formula:

$$ICC = \frac{s^2(b)}{s^2(b) + s^2(w)} \tag{5.3}$$

where $s^2(w)$ is the pooled variance within subjects, and $s^2(b)$ is the variance of the trait between subjects. Values for correlation coefficient range from -1 to 1, with 1 indicating perfect agreement, -1 total disagreement and 0 indicates that judges do not agree any more often than they disagree.

It is easily shown that $s^2(b) + s^2(w) =$ the total variance of ratings – i.e., the variance for all ratings, regardless of whether they are for the same subject or not. Hence the interpretation of the ICC as the proportion of total variance accounted for by within-subject variation.

To use the ICC with ordered-category ratings, one must assign the rating categories numeric values. In order to do that, we transformed the interpretations of each judge to numbers, so that each word was assigned 1 or 0 if it had been removed or not, respectively.

We calculated raw agreement because it is simple, intuitive and meaningful. **Proportional agreement** was calculated as the average pairwise agreement for all possible pairs

[1]Values taken from `http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm`.

of judges for a given text. We obtained three variants of it: proportion of overall agreement, proportion of positive agreement and proportion of negative agreement.

First, the pairwise **proportion of overall agreement** was calculated, that is, the proportion of cases for which each pair of judges agree for a given text, calculated as follows:

$$Proportional\ Agreement = \frac{\#\ cases\ in\ which\ judges\ agree}{\#\ cases} \tag{5.4}$$

This measure has the limitation that it can be high even if judges make judgements even by chance, most of all if one of the possible judgements is majoritary. In order to factor out agreement by chance, we calculated proportions of specific agreement and the Kappa statistic.

Proportions of specific agreement are the proportions of cases for which each pair of judges assign the same label; in our case, the proportion of cases for which judges agree on removability and the proportion for which they agree on non-removability. These proportions are interpretable as estimated conditional probabilities: for example, proportion of specific agreement on removability estimates the conditional probability, given that one of the judges, randomly selected, decides that a word has to be removed, that the other judge will take the same decision. For dicotomous judgements, as is our case, the proportions of specific agreement are calculated as follows:

$$Proportion\ of\ Specific\ Agreement(x) = \frac{2 * Ag(x)}{(2 * Ag(x)) + D} \tag{5.5}$$

where

x is one of the possible judgements
$Ag(x)$ is the number of cases in which judges agree on x
D is the number of cases in which judges disagree

The joint consideration of these two proportions of specific agreement addresses the objection that agreement may be high by chance alone, if one of the two cases considered (that words are removed or that they are not) is extremely majoritary. A high value for both proportions of specific agreement would imply that the observed level of agreement is higher than would occur by chance. Thus, by calculating both proportions, and requiring that both be high to consider agreement satisfactory, one meets the original criticism raised against raw agreement indices.

Finally, we calculated **majority agreement**, a standard index to calculate agreement when there is no gold standard. In these cases, the opinion of the majority of judges can be taken as the gold standard. In our case, any of the two judgements was considered as majoritary when it was chosen by at least half+1 of the judges that had judged the case.

In contrast with the previous measures, majority agreement is not pairwise, but it calculates the agreement of each individual judge with the majority opinion. It is calculated exactly as the proportion of overall agreement, but in this case one of the "judges" is always

the gold standard constituted by the majority. Note that the majority opinion, with which we calculate the agreement of each judge, has been calculated taking into account the opinion of that judge, which increases the expectable proportion of agreement.

$$Majority\ Agreement = \frac{\#\ cases\ in\ which\ judge\ agrees\ with\ majority}{\#\ cases} \tag{5.6}$$

The *Kappa* coefficient (Cohen 1960) has been the standard measure of stability and reproducibility of human judgements in corpus annotation since Carletta (1996). The main advantage of this measure over others, like percentage agreement between judges, is that it sorts out the possibility that judges agree by chance.

Kappa measures range from $\kappa = -1$ to $\kappa = 1$, with $\kappa = 0$ when there is no agreement other than what would be expected by chance, $\kappa = 1$ when agreement is perfect, and $\kappa = -1$ when there is systematic disagreement. Following the interpretation proposed by Krippendorf (1980) for corpus annotation, Kappa values above 0.8 indicate good stability and reproducibility of the results, while Kappa < 0.68 indicates unreliable annotation.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{5.7}$$

where

$P(A)$ is the proportion of times judges agree
$P(E)$ is the proportion of times judges are expected to agree by chance

In recent times, this measure has been criticized because the assumptions whereupon it relies are not clear enough. DiEugenio and Glass (2004) have clarified this issue, distinguishing two different computations of *kappa* that have been used indistinctively in the literature. They have highlighted the differences between the two, and hidden assumptions have been made explicit.

Less sophisticated measures of agreement do not sort out agreement by chance, but they are more transparent with respect to the aspects of the data that are taken into account, which is interesting if one wants to carry out a detailed inspection of the data. This is the reason why we also exploited correlation of judgements and raw agreement besides the kappa coefficient.

5.2 Human judgements on discourse segments

The purpose of our experiments with human judges about discourse segmentation is twofold. First, we want to check if our theoretical delimitation of discourse segment has a correspondence with the intuitions of naive judges in relation with the task of text summarization. Then, we want to gather examples that show how discourse segments are realized in text and what shallow textual clues can be found to identify them in a principled way, either manually or automatically.

Since one of the main purposes of this experiment was to check whether naive judges had a stable, reproducible concept of discourse segment, they were given only very vague, intuitive instructions about the segmentation task.

In a pilot test we had found that the concept of minimal discourse unit, or at least of discourse segment as defined by us, comes unnatural to naive judges, who feel very insecure about their own judgements, and therefore produce results of very low reliability. Taking this into consideration, we designed a more natural task, exploiting interpretative abilities that are used in everyday life, and that are related to our final task of text summarization, which is more understandable for a final user than discourse segmentation. Thus, the overt purpose of the task was to remove all non-relevant information from a text. Implicit in this task is a pre-process of segmentation: to remove fragments, one has to identify fragments first. Therefore, segmentation was performed without it being in the focus of attention of judges, which spared them insecurity in their judgements. Moreover, the kind of segmentation was targeted to text summarization, which is very adequate for our purposes.

In two words, the instructions provided to judges can be summarized as:

```
mark all intrasentential fragments of text that are not necessary to
understand the main content conveyed by the article, and that can be
removed preserving correctness of the resulting text
```

We asked that fragments of text were intrasentential because we consider sentences as the upper bound for minimal discourse units. We were aware that this restriction might make the task unnatural, but judges did not mention it as a problem in the questionnaires that were passed to them after the segmentation task. Moreover, multiple sentences could be removed by removing all the words in each sentence.

Our starting hypotheses are the following:

Hypothesis 1

the probability that a word is removed is not homogeneous, but some words have a higher probability to be removed than others.

null hypothesis *the probability that a word is removed is homogeneous across text.*

Hypothesis 2

the probability that a word is removed is conditioned by the fact that it belongs to a certain latent class.

Hypothesis 2a

words have different probabilities of being removed conditioned by the fact that they belong to a **marked** *or an* **unmarked** *segment.*

Hypothesis 2b

words have different probabilities of being removed conditioned by the fact that they are marked by **different textual cues**.

Hypothesis 2c

words have different probabilities of being removed conditioned by the fact that they belong to segments with different **discourse semantics**.

null hypothesis *the probability that a word is removed is homogeneous across classes.*

Each of these hypotheses is tested against its corresponding null hypothesis, in Sections 5.2.3 and 5.2.4. If the null hypotheses are rejected, it can be assumed that annotations by human judges support our theoretical proposal about the organization of text at discourse level. In order to test the two main hypotheses, we have taken a *latent class model* approach.

The premise of a latent class model approach is that cases in a population belong to two or more classes, and that these classes are latent because the class membership of a given case is not directly observed. A case's probability of being assigned a given label is assumed to depend on the case's latent class. In our case, this means that the probability of a word to be removed dependens on the class it belongs to, for example, if it belongs to the class of "*words in a relative clause*" or to the class of "*words in Named Entity*". The aims of the analysis are to estimate, for each latent class (*relative*, *punctuation*, *cause*, *equality*, etc.), the probability that a member of that latent class will be assigned a given label from a judge.

If the probability of a word being removed by a judge is conditioned by its latent class, that is, if it differs if the word belongs to different classes, then we can consider that the null hypothesis is rejected.

In order to test these hypotheses against their corresponding null hypotheses, we carry out hypotheses tests, but we also obtain a number of measures of agreement between judges: correlation between the judgements of different judges, inclusive and exclusive majority agreement, overall agreement and specific (positive and negative) agreement, and the standard *kappa* coefficient. We believe that if judges agree significantly above what would be expected by chance, the null hypothesis can be rejected.

	n. of words	n. of sentences	n. of segment boundaries
apa 1	316	17	52
apa 2	1194	47	165
apa 3	235	8	28
apa 4	649	21	72
apa 5	671	31	75
text 1	354	12	55
text 2	182	5	25
text 3	797	30	115
text 4	151	3	12
text 5	741	17	130
text 6	855	37	102
text 7	661	23	109
text 8	255	8	45
text 9	593	20	83
total	**7640**	269	1068

Table 5.2: Size of the newspaper articles that were manually segmented by naive judges, and their distribution in terms of linguistic units.

5.2.1 Corpus and judges

35 judges carried out the subsentential reduction of a corpus of 14 newspaper articles dealing with various topics: finance, international and local news, critic of cinema, etc. The corpus contains a total of 7640 words in 269 sentences, the sizes and distribution of units through the corpus can be seen in Table 5.2. Each article was reduced by at least 4 judges.

Judges were not trained for the task, in most cases, they were not even linguists, although most of them had some kind of universitary education. They carried out their judgements in paper or electronic support, depending on their familiarity with electronic support. They marked those spans of text (or isolated words) that they considered removable. Judgements were transformed to the standard format shown in Figure 5.2. In this form it can be readily seen that the agreement of judges on removability increases nearby boundaries, specially when they are marked by punctuation.

In order to carry out a latent class analysis as required to test the hypotheses 2a, 2b and 2c above, different subsets of the texts were manually annotated so that each word was assigned to one of three different inventories of latent classes:

– discourse segments were identified following the definition provided in Section 3.3.2, making a difference between *marked* and *unmarked* segments. As explained in Section 3.3.2.2, the argumental core of a clause can be considered as the unmarked discourse segment, while marked discourse segments are dominated by discourse markers, syntactical structures, etc.

JERUSALEN.-	X	punctuation
La		
muerte		
de		
cinco		
policías		
palestinos		
abatidos	X	participle
por		
el		
Ejército		
Israelí		
cerca	X	discourse marker, adjunct
de	X	
Ramala	X	
(Cisjordania)	XX	punctuation, adjunct
durante	XX	discourse marker, adjunct
la	XX	
noche	XX	
del	XX	
domingo	XX	
al	XXX	adjunct
lunes	XXX	
pasado	XX	
fue		
un		
"error"		
debido	X	discourse marker
a	X	
"malas	X	
informaciones"	X	
,	X	punctuation
según	X	discourse marker
ha	XX X	reporting construction
señalado	XX X	
un	X	
alto	X	
responsable	X	
israelí	X	
que	XXXX	relative clause
ha	XXXX	
pedido	XXXX	
el	XXXX	
anonimato	XXXX	
.		punctuation
El		
Ejército		
israelí		
ha		reporting construction
propuesto		
iniciar		
una		
investigación		
al		
respecto		
.		punctuation
"Los		
soldados		
pensaban		reporting construction
que		
atacaban		
una		
posición		
ocupada		participle
por		
miembros		
de		
la	X	
Fuerza	X	
17	X	
(la	X	punctuation, apposition
guardia	X	
personal	X	
del	X	
presidente	X X	
palestino	X X	
Yasir	X	apposition
Arafat)	X	
o	X	discourse marker
de	XX	
otro	XX	
servicio	X	
de	X	
seguridad	X	
que	X	relative clause
,	XXX	punctuation, discourse marker, adjunct
desde	XXX	
hace	XXX	
dos	XXX	
semanas	XXX	
,	XXX	

dispara	X	
diariamente	XX	adjunct
contra	X	discourse marker
un	X	
campamento	X	
militar	X	
israelí	XX	
y	XX	discourse marker
una	XX	
carretera	XX	
estratégica	XX	
situada	XX	
en	XX	
las	XX	
proximidades"	XX	
,	XXX	punctuation, reporting construction
ha	XXX	
indicado	XXX	
este	XXX	
responsable	XXX	
.		
"Los	X X	punctuation
militares	X X	
no	X X	
sabían	X X	reporting construction
que	X X	
los	X X	
miembros	X X	
de	X X	
la	X X	
Fuerza	X X	
17	X X	
habían	X X	
sido	X X	
reemplazados	X X	
por	X X	
policías	X X	
que	XX X	relative clause
no	XX X	
se	XX X	
dedicaban	XX X	
a	XX X	
actividades	XX X	
terroristas"	XX X	
,	XXXX	punctuation, reporting construction
ha	XXXX	
añadido	XXXX	
.		
Según	XXXX	discourse marker, punctuation, adjunct
los	XXXX	
medios	XXXX	
de	XXXX	
comunicación	XXXX	
,	XXXX	
el	X	
jefe	X	
del	X	
Estado	X	
Mayor	X	
Israelí	X	
,	XXXX	punctuation, apposition
el	X XX	
general	X XX	
Saul	X XX	
Mofaz	X XX	
,	X XX	
propuso		
constituir		
una		
comisión		
de		
investigación		
sobre		
este		
incidente		
durante	XX X	discourse marker, adjunct
una	XX X	
comparecencia	XX X	
a	XX X	
puerta	XX X	
cerrada	XX X	
ante	XX X	discourse marker, adjunct
la	XX X	
comisión	XX X	
de	XX X	
Defensa	XX X	
y	XX X	
Asuntos	XX X	
Exteriores	XX X	
del	XX X	
Parlamento	XX X	

Figure 5.2: Example text segmented by human judges. Columns are judges, rows are words in the text. Crosses indicate that the judge in that column considered the word in that row as removable. Potential segment boundaries are marked, irrespective of whether

- for 5 of the articles (*apa1* to *apa5*), discourse relations were manually annotated, as described in Section 5.3; segments were classified by the semantics of this relation.
- all textual evidence that was liable to mark an intrasentential segment boundary was marked. The items marked as evidence of a segment boundary outnumber the actual segments in the text, because a single segment can be marked by more than one linguistic feature, as in the following example, where the segment "*, promoviendo una idea que sabe molesta para muchos colegas*" is marked both by punctuation and an absolutive participle.

(1) `Ahora está en plena cruzada, promoviendo una idea que sabe molesta para muchos colegas:`

	Ahora está en plena cruzada, promoviendo una idea que sabe molesta para muchos colegas:
punctuation	, promoviendo una idea que sabe molesta para muchos colegas:
participle	, promoviendo una idea que sabe molesta para muchos colegas:
context	, promoviendo una idea que sabe molesta para muchos colegas:
relative	que sabe molesta para muchos colegas:
discourse marker	para muchos colegas:
context	para muchos colegas:

As detailed in Table 5.2, in the part of the corpus that was tagged with this information, we identified a total of and a total of 1068 items as textual evidences of segment boundaries, which implies a mean of 8.31 per sentence. The distribution of segments, of different kinds of segments (distinguished by form and by semantics) and the total number of boundaries and the size of the segments under their scope can be seen in Tables 5.3 and 5.4.

(2) El problema, en su opinión, es que una teoría del todo no podrá explicar jamás algunos de los problemas más importantes de la ciencia de hoy, que tienen que ver con la aparición de propiedades nuevas en sistemas construidos con un número elevado de partículas.

Note that markedness does not necessarily imply a higher probabilty that the word is removed. A marked discourse segment can be accessory to the main content conveyed in a text, as in the previous example, but it can also provide crucial information, as in the following example. In this example, the underlined text is a marked segment that introduces an element that is crucial to understand the implications of the sentence.

(3) 'Gran parte de la física moderna se basa en creencias reduccionistas, más que en hechos experimentales, y esto será muy perjudicial para la ciencia a largo plazo'

Even despite the lack of correspondence between markedness and probability of being removed, we believe that it is useful to study the behaviour of judges as conditioned by the markedness of segments. In contrast, we believe that marking segments in text by their removability may introduce bias in the judgement. Indeed, identifying "removable" segments

	n. of marked segments	n. of words in marked segments	average marked segment length
apa 1	33	206 (65%)	6.2
apa 2	108	806 (67%)	7.4
apa 3	20	164 (70%)	8.2
apa 4	47	426 (65%)	9.0
apa 5	52	380 (56%)	7.3
text 1	36	201 (56%)	5.6
text 2	16	126 (69%)	7.9
text 3	73	400 (50%)	5.5
text 4	18	123 (81%)	6.8
text 5	75	509 (68%)	6.8
text 6	61	399 (46%)	6.5
text 7	48	260 (39%)	5.4
text 8	25	142 (55%)	5.7
text 9	51	306 (51%)	6.0

Table 5.3: Distribution of marked and unmarked segment in the texts manually segmented by naive judges.

is more subjective than identifying marked ones, and so it is very probable that the text that we would take as a reference for latent classes is unstable itself. As a consequence, the comparison with judges may be error-prone.

The distribution of marked segments in the corpus can be seen in Table 5.3. The kind of evidence that we have considered to distinguish segments by their form was:

punctuation words under the scope of a segment delimited by punctuation boundaries.

(4) Según los medios de comunicación, el jefe del Estado Mayor israelí [,] el general Saul Mofaz [,] propuso constituir una comisión de investigación sobre este incidente durante una comparecencia a puerta cerrada ante la comisión de Defensa y Asuntos Exteriores del Parlamento, dando a entender claramente también que se había cometido un error.

parentheses words under the scope of a segment delimited by parenthetical punctuation.

(5) Además, como no hay una única regla que describa estos sistemas [–] como desearía el ideal reduccionista [–], sólo es posible entender su funcionamiento 'mediante la observación, y logrando que teoría y experimento se den la mano', afirma Laughlin.

reporting speech constructions words belonging to a construction that serves to introduce something that has been said, probably by another person.

(6) "Los militares no sabían que los miembros de la Fuerza 17 habían sido reemplazados por policías que no se dedicaban a actividades terroristas", ha añadido.

participle phrase words belonging to a phrase headed by a participle (past or present, also known as gerund) that is not creating a composite verb together with an auxiliary verb.

(7) Según los medios de comunicación, el jefe del Estado Mayor israelí, el general Saul Mofaz, propuso constituir una comisión de investigación sobre este incidente durante una comparecencia a puerta cerrada ante la comisión de Defensa y Asuntos Exteriores del Parlamento, dando a entender claramente también que se había cometido un error.

relative clause words belonging to clauses subordinated by a relative pronoun.

(8) "Los militares no sabían que los miembros de la Fuerza 17 habían sido reemplazados por policías que no se dedicaban a actividades terroristas"

apposition words in a segment that rewords the information conveyed by the immediately previous segment, typically, an entity.

(9) Las palabras de Arafat, que en aquel momento se encontraba en Egipto, fueron contestadas de manera automática desde Jerusalén por el ministro de Asuntos Exteriores israelí, Simón Peres.

highly grammaticalized discourse markers words in a segment dominated by a highly grammaticalized discourse marker, which are usually ambiguous between sentential and discursive function.

(10) él explicó lo que estaba pasando y los tres se llevaron el Premio Nobel de Física en 1998.

discourse marker words in a segment dominated by a discourse marker

(11) La muerte de cinco policías palestinos abatidos por el Ejército israelí cerca de Ramala (Cisjordania) durante la noche del domingo al lunes pasado fue un "error" debido a "malas informaciones", según ha señalado un alto responsable israelí que ha pedido el anonimato.

Each of these segments was distributed in the corpus (texts 1 to 9) as can be seen in Table 5.4.

type of segment	n. of segments	n. of words dominated by segment	average segment length
segments by form			
punctuation	363	3249	8.9
parentheses	35	210	6.0
reporting	46	213	4.6
participle	49	358	7.3
relative	122	1531	12.5
apposition	79	801	10.1
discourse marker	287	2925	10.1
grammatical disc. mark.	96	1094	11.3
segments by semantics			
symmetric	118	1941	16.45
asymmetric	364	2266	6.22
continuation	209	2178	10.42
elaboration	273	2009	7.36
context	122	695	5.70
parallelism (*equality)*	106	1561	14.73
cause	37	356	9.62
revision	16	213	13.31

Table 5.4: Distribution of segment types in the texts manually segmented by naive judges.

5.2.2 Study of variance

We studied the variance in the task of removing irrelevant words from text, a good indicator of the stability of the annotation, which is a necessary requirement for results to be reproducible.

Figure 5.3 displays the variance of words with respect to their probability of being removed, calculated as the proportion of judges that actually removed a given word. It can be seen that the behaviour of judges in removing words is not homogeneous, because the boxplot has long tails, and the standard deviation is rather high: $sdev = .326$. This high variance seems to indicate that words may belong to different classes, because judges have significantly different behaviours with respect to different words. Our target is to model the underlying reasons for differences accross words, distinguishing different classes of words which present a lower internal variance of the behaviour of judges with respect to its removability.

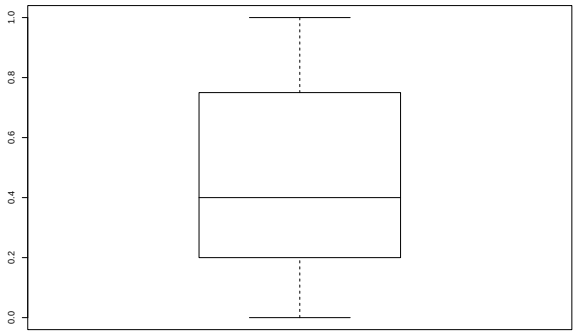


Figure 5.3: Distribution of the different values of the probability that a word is removed.

In the first place, we need to test whether the behaviour of judges with respect to words is comparable enough to be exploited as the basis to infer classes of words and test their descriptive adequacy. To do that, we will first explore the variability of judges with respect to the ratio of reduction of texts by removing words. In Figure 5.4 we can see the distribution of the different values of the ratio of words that each judge removed in each text.

We can see that most texts present a distribution of values close to normal, thus indicating that most judges behave in a comparable way. Some texts, like apa1, apa2, apa3, text3 or text6 have outliers, but this does not seem to strongly affect their standard deviation, thus indicating that the presence of outliers may be disregarded as a source of noise in obtaining classes of words. However, a big variance in the distribution of values, as in text9, seems to affect the use of the data for inducing models, as will be shown when we apply hypothesis testing, in the next section.

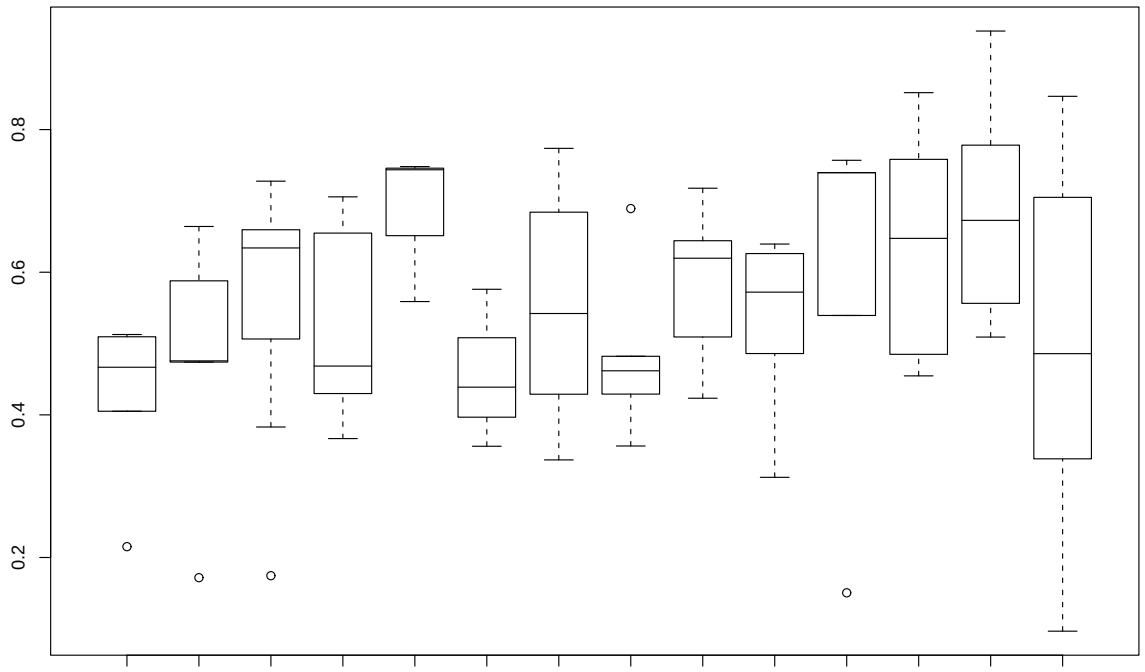


Figure 5.4: Graphical representation of distribution of the values for the ratio of words removed by each judged from a given text (from left to right: apa1, apa2, apa3, apa4, apa5, text1, text2, text3, text4, text5, text6, text7, text8, text9).

5.2.3 Probability that a word is removed

Hypothesis 1

the probability that a word is removed is not homogeneous, but some words have a higher probability to be removed than others.

null hypothesis *the probability that a word is removed is homogeneous across text.*

The probability that a word is removed is the proportion of judges that actually removed that word in a text with respect to all the judges that reduced a given text. We obtained that the overall probability that any word in this corpus is removed is .46, with a high variance among probabilities for different words (with a standard deviation of .32), which suggests that partitioning words may have different behaviours with respect to relevance according to the class they belong to.

In order to test this hypothesis, we need to find systematicities in the way judges

text	t	p-value	degrees of freedom (n. of words)	n. judges	κ	agreement
apa1	-1.9	.05	316	6	.34	.67
apa2	-1.9	.04	1192	5	.17	57
apa3	-1.8	.06	233	7	.23	.61
apa4	-2.0	.04	647	5	.38	.68
apa5	-9.1	$< .005$	669	3	.31	.69
text1	-1.8	.06	336	4	.41	.70
text2	-1.7	.09	188	8	.26	.63
text3	-1.3	.20	809	6	.20	.59
text4	-2.0	.04	161	5	.34	.67
text5	-1.9	.05	747	6	.15	.57
text6	-6.3	$< .005$	862	5	.26	.60
text7	-8.0	$< .005$	693	5	.14	.59
text8	-7.1	$< .005$	273	11	.28	.69
text9	-0.1	.95	598	8	.29	.63

Table 5.5: Results of applying hypothesis testing to the actual probability that a word is removed against a probability produced at random.

remove words from text. If we cannot find any organization in the judgements, then we cannot use the data to support a theory of discourse organization. In order to find these systematicities, we first test whether the judgements are different from random. Then, we assess to what extent judges agree with each other. Finally, we combine these two approaches to get a better insight on the data.

5.2.3.1 Hypothesis testing

We want to test whether the actual distribution of probabilities of words being removed could have been produced by random or else it is conditioned to some underlying organization of text. In order to do that, for each word in a text, we compared the actual probability that it was removed with a probability produced randomly.

As can be seen in Table 5.5, the results obtained from applying the matched samples t-test provide enough evidence to reject the null hypothesis, although not strongly, because significance values are usually not significantly below $p = .05$, and in some cases they are well above (text3, text9). Nevertheless, the value of t is almost always beyond the critical value, which is $t_{crit} = \pm 1.645$ for $p = .05$, as described in Table 5.1. It is noteworthy that when results are more reliable, that is, for lower values of p, the value of p increases significantly (apa5, text6, text7, text8). It must also be observed that significance drops dramatically when the judgements for a given text differ, as is the case of text9, discussed in the previous section.

These results indicate that the behaviour of judges differs from random, although it is still rather chaotic. More significant conclusions could be obtained from a bigger corpus.

majority opinion	overall agreement	agreement on removability	agreement on non-removability	Kappa coefficient
0.80	0.68	0.57	0.72	0.26

Table 5.6: Different measures of agreement between judges for removability of words.

judges	majority opinion	overall agreement	agreement on removability	agreement on non-removability	Kappa coefficient
o1	.82	.65	.53	.70	.27
e2	.89	.79	.69	.84	.54
l2	.83	.68	.72	.61	.34
n1	.86	.72	.63	.73	.44
e1	.76	.53	.51	.55	.08

Table 5.7: Measures of agreement in judging the removability of words of the same texts between the same judge and herself two years later.

5.2.3.2 Agreement between judges

Besides providing an insight about the behaviour of judges, measures of agreement between judges are good indicators of the reproducibility of the annotation.

In the first place, we applied the standard measures of agreement between judges described in Section 5.1. As can be seen in Table 5.6, the agreement between judges is rather low. The value of Kappa even if it is not chance agreement (0), it is far below the threshold that indicates stable and reproducible results (.68, Carletta 1996). The proportion of agreement (overall agreement) and specific agreements (positive and negative agreement, agreement on removability and on non-removability, respectively[2]) are also close to what could be expected by chance.

We also carried an experiment, asking some judges to remove words from the same texts that they had already reduced one year later. The measures for agreement of a judge with herself a year later can be seen in Table 5.7. It can be seen that, even if some judges show a higher agreement than the average agreement between different judges, they are still far from reaching the levels that indicate stability.

The agreement of judges with respect to the words to be eliminated is higher than could be expected by chance, but it is still far from values that would indicate good reproducibility of results. However, if we analyze the patterns of behaviour of judges in detail, it can be seen that groups of judges can be distinguished according to their behaviour. As can be expected, the agreement between judges is higher for judges with the same behaviour.

In Table 5.9 we display the agreement between judges by giving the correlation coefficient between the judgements of different judges for the same text. The cases where significance values were above $p = 0.005$ have been marked in boldface. It can be seen that

[2]proportion of positive agreement: agreement on removability, that is, the proportion of words in which judges agree that they have to be removed; proportion of negative agreement: agreement on non-removability, that is, the proportion of words in which judges agree that they do not have to be removed

lower values of significance correspond to judges with lower correlation (that is, with lower agreement) with other judges.

Interestingly, cases of low correlation are not isolated, but it tends to happen that judges with low correlation with some other judge present low correlations with most of the judges in a given text. In many cases, the highest correlations of low-correlation judges occur with the other judges that also have low correlations in the same text, as can be seen by the boxed figures in text 2 and 8.

In order to test whether it makes sense to separate judges in groups, as if they had the same pattern of behaviour, we separated the judges in text 8 in two groups: high-correlation (judges e2, g1, i1, l2, n1, o1, s1, v1) and low-correlation (judges j1, m1, m2). We applied the t-test and measures of agreement for each group separatedly, results can be seen in Table 5.8. As could be expected, the measures of agreement were higher for high-correlation judges, but the value for the t-test was higher for low-correlation judges.

text	t	p-value	degrees of freedom (n. of words)	n. judges	κ	agreement
text8 - total	-7.1	$< .005$	273	11	.28	.69
text8 - high corr.	-4.4	$> .005$	273	8	.39	.72
text8 - low corr.	-11.5	$> .005$	273	3	.14	.69

Table 5.8: Results of applying hypothesis testing to subgroups high- and low-correlation judges, after the correlations obtained from Table 5.9.

It seems clear that separating judges into groups with homogeneous behaviours may be useful to reduce noise in the data and obtain a more accurate model. However, in the following experiments we have not separated judges, because we believe that there is little evidence to carry out such separation with enough reliability, even in texts like 8, with a lot of judges and big differences between them.

text 3

	a3	d2	j2	m4	v2
a1	.10	.10	**.04**	.17	.12
a3		.27	.21	.23	.64
d2			**-.13**	.39	.20
j2				.26	.24
m4					.24

text 5

	e1	g1	l1	m5	v2
d2	.26	.10	.18	.25	.28
e1		**.03**	.10	.25	.31
g1			**-.03**	**.04**	.25
l1				.10	.12
m5					.14

apa 1

	d3	d4	g3	i1	n1
b2	.47	.52	.20	.34	.52
d3		.40	**.13**	.43	.55
d4			.40	.20	.27
g3				**.03**	**.12**
i1					.81

apa 3

	b2	c1	d3	d4	g3	l3
a5	.20	.35	.44	.32	.18	.46
b2		.28	**-.11**	.51	.32	.19
c1			.31	.23	**.16**	.32
d3				.20	**.08**	.34
d4					.22	.40
g3						.36

text 2

	g2	i1	l2	m4	m6	n1	o1
e1	.53	**-.07**	.34	.78	**-.06**	.23	.27
g2		**.06**	**.08**	.54	**-.15**	.43	.52
i1			.34	**.17**	.48	.26	.38
l2				.52	**.14**	.35	.45
m4					**-.04**	.30	.41
m6						**.12**	**.14**
n1							.50

text 9

	e2	l2	m1	m3	o1	s1	v1
b1	.42	.61	.27	.35	.30	.49	.53
e2		.53	.35	.31	.55	.28	.57
l2			.24	.31	.49	.34	.42
m1				.42	**.04**	.12	.29
m3					.25	.18	.48
o1						.23	.46
s1							.31

text 8

	g1	i1	j1	l2	m1	m2	n1	o1	s1	v1
e2	.26	.31	**.16**	.73	**-.01**	.35	.44	.61	.72	.69
g1		.35	**.02**	**.15**	.27	.44	.20	**-.09**	.49	.23
i1			**-.03**	.31	**.16**	.51	.53	.19	.39	.37
j1				**.03**	.40	**-.05**	**.11**	.31	**.04**	**-.01**
l2					**.01**	.35	.39	.49	.61	.62
m1						.30	**.08**	**.11**	**-.01**	**-.01**
m2							.31	.19	.44	.45
n1								.45	.55	.49
o1									.40	.40
s1										.67

Table 5.9: Correlation of judgements for each pair of judges that have judged the same text (considering only texts judged by more than 5 people), each judge is identified by a unique combination of a number and a letter.

5.2.4 Latent class analysis

Hypothesis 2

the probability that a word is removed is conditioned by the fact that it belongs to a certain latent class.

Hypothesis 2a

words have different probabilities of being removed conditioned by the fact that they belong to a **marked** *or an* **unmarked** *segment.*

Hypothesis 2b

words have different probabilities of being removed conditioned by the fact that they are marked by **different textual cues**.

Hypothesis 2c

words have different probabilities of being removed conditioned by the fact that they belong to segments with different **discourse semantics**.

null hypothesis *the probability that a word is removed is homogeneous across classes.*

We assume that each word can be classified into a *latent class* that explains its role with respect to keeping the relevance and coherence of the text in summarization. In our case, we explore how well the classes motivated theoretically in Chapter 4 can group words with a homogeneous role, while distinguishing heterogeneous words. We understand that homogeneity in the role of words is reflected in a homogeneous behaviour of judges with respect to these words, so that both the probability that a word is removed and inter-judge agreement are different across classes. We have explored three kinds of theoretically motivated classes:

1. *marked* vs. *unmarked* segments,
2. segments distinguished by their surface form,
3. segments distinguished by their discourse semantics, and
4. segments distinguished by their surface form **and** their discourse semantics.

5.2.4.1 Marked vs. unmarked segments

In this section we check whether the probability that a word is removed is affected by the fact that it occurs in a marked vs. an unmarked segment, in other words, whether the human judgements about the relevance of words can be modelled considering the degrees of markedness of discourse segments as latent classes. As explained in Section 3.3.2.2, typical unmarked discourse segments are the argumental core of clauses, while marked discourse

	average	a1	a2	a3	a4	a5	t10	t2	t3	t4	t5	t6	t7	t8	t9
total	**.44**	.56	.52	.45	.46	.30	.55	.44	.51	.40	.46	.41	.35	.30	.49
marked	**.54**	.61	.57	.51	.57	.37	.74	.52	.65	.49	.52	.55	.40	.46	.67
unmarked	**.26**	.36	.34	.22	.18	.17	.30	.25	.36	.16	.32	.29	.32	.12	.30

Table 5.10: Ratio of reduction of words, specifying the ratios for words belonging to *marked* and *unmarked* segments.

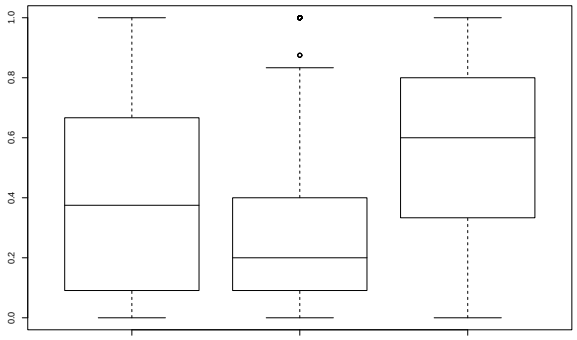


Figure 5.5: Distribution of the probabilities that a word is removed, across all words (leftmost box, *sdev* = .326) and words belonging to *unmarked* (center, *sdev* = .296) and *marked* (rightmost, *sdev* = .298) segments.

segments are dominated by marked syntactical structures (e.g., hypotactic constructions), lexical pieces (e.g., discourse markers) or prosodic positions (e.g., dislocated to the left).

To begin with, the probability that a word is removed is slightly different in marked and unmarked segments. As can be seen in Table 5.10, words within a marked segment have a probability of .54 of being removed, whereas the words in unmarked segments have a probability of .26. This indicates that a property of the latent class *marked* is that words belonging to this class are perceived as less relevant than words belonging to the class of *unmarked*.

It is also interesting to note that the classes of *marked* and *unmarked* are slightly more homogeneous with respect to the perception of relevance than considering all classes indistinguishably, as can be seen in the boxplots in Figure 5.5. The length of the boxes is proportional to the variability in the probability that a word is removed: across all words, the standard deviation of this probability is *sdev* = .326, while it is *sdev* = .249 within the class of *unmarked* words and *sdev* = .303 within *marked*. In all cases, the variability is high, but the fact that it is smaller within classes indicates that dividing words into latent classes is useful to group homogeneous words.

As follows, latent classes are also useful to distinguish heterogeneous words. The classes of marked and unmarked segments can be considered heterogeneous because the words they contain are perceived differently by judges: words in marked segments have a higher

	majority opinion	overall agreement	specific removability	specific non-removability	Kappa
total	.80	.68	.57	.72	.26
marked	.75	.61	.62	.53	.22
unmarked	.79	.67	.33	.75	.27

Table 5.11: Inter-judge agreement measures for removability, distinguishing *marked* and *unmarked* discourse segments.

probability of being removed. Indeed, we obtained $t = -39.24$ (with $p < .0005$) for a *t-test* comparing the distribution of the probabilities of being removed for words in marked and unmarked segments, thus rejecting the null hypothesis that the two distributions are equal. The value for t is smaller when comparing either marked or unmarked to all the words without distinction ($t = -26.10$ for marked, $t = -17.77$ for unmarked).

In addition, we found that measures of agreement between judges are also sensitive to the classification of words into *marked* and *unmarked*. As seen in Table 5.11, judges tend to agree about the relevance of words more in unmarked segments than in marked ones, although with only small differences.

However, if we take a closer look to the values for *specific agreements*, we will see there is a difference of 42 points between positive (.33) and negative (.75) specific agreement for unmarked segments, whereas the difference for marked segments is of only 9 points. This seems to confirm the idea that words in marked and unmarked segments are perceived differently. Indeed, the majority of judges agree on which are the relevant (that is, non-removable) words in unmarked segments, but there is much disagreement on which words are removable, and, within marked segments, there is also disagreement on which words are relevant.

From these results we can conclude that the classes of words belonging to marked and unmarked segments are indeed distinct. However, the high variability in the perception of relevance by judges within these classes, especially within the class of words belonging to a marked segment, indicates these are not the not optimal latent classes for the task of summarization. In the following sections we will subdivide these classes in order to find an organization of the data that provides a more adequate description of the perception of relevance by judges.

5.2.4.2 Discourse segments typified by surface form

In this section we check whether the probability that a word is removed depends on the surface form of the segment where it occurs, in other words, whether judgements about the relevance of words depend on the form of its containing discourse segment. This implies a division of the class of words belonging to *marked* segments in subclasses, while the class of *unmarked* remains united as a whole.

Let us recall that we have manually identified the following kinds of discourse segment:

	average	a1	a2	a3	a4	a5	t10	t2	t3	t4	t5	t6	t7	t8	t9
total	**.44**	.56	.52	.45	.46	.30	.55	.44	.51	.40	.46	.41	.33	.30	.49
punctuation	**.58**	.71	.62	.68	.65	.45	.70	.63	.67	.51	.49	.61	.43	.38	.60
parentheses	**.78**	.97	.92		1	1	.52	.75	.83	.62	.83	.08	.60	.63	
reporting	**.62**	.59	.63			.66	.74	.46	.77	.40	.58	.77	.55	.86	.47
participle	**.60**	.66	.66	.74	.87		.53	.54	.55		.47	.74	.50	.30	.69
relative	**.58**	1	.60	.64	.73	.41	.81	.65	.69	.20	.51	.58	.46	.34	.62
apposition	**.74**	.97	.76	.81	.91	.93	.91	.72	.82	.64	.39	.79	.58	.35	.85
disc. mark.	**.51**	.56	.58	.52	.53	.37	.75	.58	.62	.41	.44	.49	.37	.42	.56
gram. disc.	**.49**	.60	.62	.61	.68	.28	.55	.51	.55		.56	.46	.34	.27	.42

Table 5.12: Ratio of reduction of discourse segments distinguished by their form.

- punctuation
- parentheses
- reporting speech
- participles
- relatives
- appositions
- highly grammaticalized discourse markers
- discourse markers

As can be seen in Table 5.12, the probability that a word is removed is significantly different across classes. In some classes (parentheses, apposition) the probability that a word is removed is high, in some others (punctuation, reporting, participle, relative) it is slightly higher than 50%, and in the classes of words dominated by a discourse marker or a highly grammaticalized discourse marker it is very close to the mean taking all word classes indistinctly. Interestingly, the probability that a word ocurring in discourse segment marked formally is removed is always higher than the average.

In Figure 5.6 we can see the variability in the probability that a word is removed within these classes of words. In all the cases, the variability in classes of words marked by a surface clue is smaller than the mean taking all words indistinguishably (leftmost box), so words grouped in these classes are more homogeneous. However, the high intra-class variability shows that heterogeneous words are not properly distinguished.

As can be seen in Table 5.13, measures of agreement between judges, especially the kappa coefficient, reflect the heterogeneity of the words contained in different classes. Classes with homogeneous words present higher levels of agreement (parentheses, reporting, relative, apposition), while classes grouping heterogeneous words present lower agreement, even below the average agreement for all words indistinguishably (discourse marker, punctuation, participle). It has to be noted that the classes with highest agreement, *parentheses* and *reporting* present a very limited coverage of the texts.

These results indicate that dividing words into classes determined by their form provides a model of how judges perceive relevance that is better than the null hypothesis, but still far from optimal.

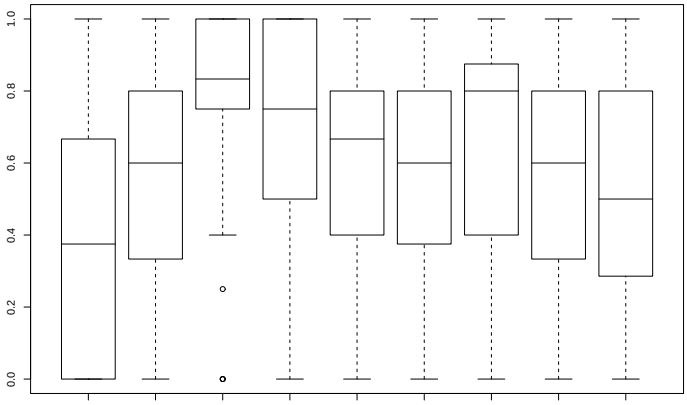


Figure 5.6: Distribution of the probabilities that a word is removed within segments distinguished by their form, from left to right: all words indistinguishably ($sdev = .326$), segments marked by punctuation ($sdev = .307$), parentheses ($sdev = .190$), reporting ($sdev = .304$), participle ($sdev = .265$), relative ($sdev = .292$), apposition ($sdev = .301$), a highly grammaticalized discourse marker ($sdev = .294$) or a discourse marker ($sdev = .316$).

	majority opinion	overall agreement	specific removability	specific non-removability	Kappa
total	.80	.68	.57	.72	.26
punctuation	.77	.63	.65	.50	.23
parentheses	.81	.69	.68	.10	.77
reporting	.73	.59	.56	.34	.51
participle	.75	.61	.61	.40	.25
relative	.78	.64	.61	.44	.32
apposition	.83	.72	.73	.25	.38
discourse marker	.76	.62	.60	.55	.24
gram. disc. mark.	.76	.62	.56	.56	.28

Table 5.13: Inter-judge agreement measures for removability, distinguishing segments by their form.

	average	a1	a2	a3	a4	a5
total	**.40**	.56	.44	.33	.38	.30
symmetric	**.41**	.53	.47	.43	.36	.30
asymmetric	**.43**	.57	.46	.44	.41	.27
elaboration	**.48**	.64	.53	.46	.47	.33
continuation	**.41**	.53	.45	.41	.39	.29
context	**.47**	.72	.56	.32	.49	.30
equality	**.41**	.55	.47	.42	.36	.29
cause	**.27**		.46	.03	.35	.27
revision	**.56**	1.00	.50	.50	.46	.36

Table 5.14: Ratio of reduction of discourse segments of different semantics.

5.2.4.3 Discourse segments typified by their semantics

In this section we check whether the probability that a word is removed depends on the discourse semantics of the segment where it occurs, in other words, whether judgements about the relevance of words depend on whether their containing segment is conveying a causal relation, revision, equality, etc.

In order to determine the semantics of discourse segments, we have used the corpus presented in Section 5.3, where three independent judges identified relations between discourse segments and labelled their semantics. To create a corpus annotated with semantic features, segments were labelled with those semantic features for which at least 2 judges agreed. A more detailed description of the list of semantic features can be found in Table 5.18. Note that the set of features used in this annotation has two features (*symmetric* and *asymmetric*) that have not been included in the final inventory explained in Chapter 4 for reasons that will be explained in the following section.

As can be seen in Table 5.14, the probability that a word is removed does not significantly differ across classes. Words occurring in segments of *cause* have 13% lower probability of being removed than the average, and words occurring in segments of *revision* have a 16% higher. *Context* and *elaboration* present probabilities slightly higher than the average. This indicates that the semantics of segments does not seem to be a good indicator of how the relevance of words is perceived.

In Figure 5.7 we can see that the variability in the probability that a word is removed is high within classes, in some cases, it is even higher than considering all words indistinguishably (asymmetric, elaboration, continuation, equality, cause). Therefore, these classes seem to contain heterogeneous kinds of words.

However, as seen in Table 5.15, the kappa coefficient of inter-judge agreement is rather high for the marked cases for the *semantic* dimension of discursive meaning (*cause*, *revision* and *equality*, see Section 4.2.3.1.2), while the least marked case, *context*, presents lower agreement than considering all words indistinguishably. Also the most marked case in the *structural* dimension, *continuation* has higher agreement than the least marked, *elaboration*.

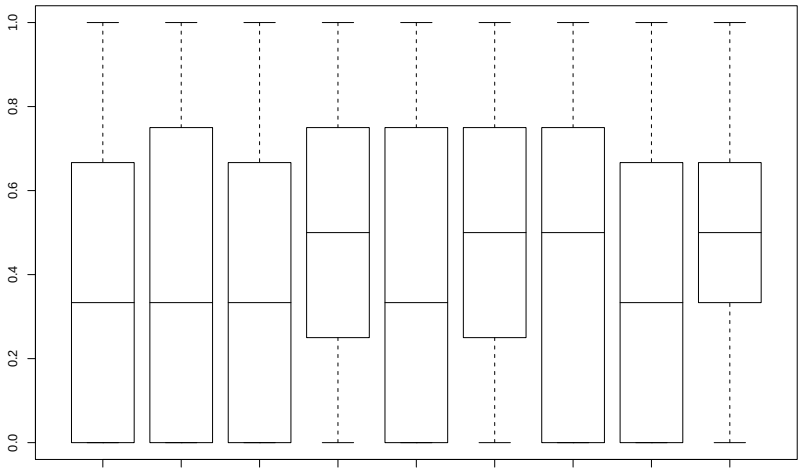


Figure 5.7: Distribution of the probabilities that a word is removed within segments distinguished by their discursive semantics, left to right: all words ($sdev$ = .348), symmetric ($sdev$ = .343), asymmetric ($sdev$ = .350), elaboration ($sdev$ = .353), continuation ($sdev$ = .351), context ($sdev$ = .340), equality ($sdev$ = .358), cause ($sdev$ = .355), revision ($sdev$ = .296).

	majority opinion	overall agreement	specific removability	specific non-removability	Kappa
average	.82	.70	.56	.76	.28
symmetric	.78	.65	.57	.67	.28
asymmetric	.79	.65	.58	.68	.28
elaboration	.78	.64	.62	.63	.27
continuation	.80	.67	.59	.70	.32
context	.77	.63	.58	.58	.20
equality	.81	.69	.63	.71	.37
cause	.87	.76	.43	.81	.53
revision	.73	.57	.47	.34	.48

Table 5.15: Inter-judge agreement measures for removability of segments of different semantics.

5.2.4.4 Discourse segments typified by their semantics and their form

In the two previous sections we have tried to distinguish subclasses within the class of words within *marked* discourse segments. We have found that the form of discourse segments is useful to distinguish words with heterogeneous behaviour, but only in highly marked cases with very little coverage of real corpora. We have found that the *semantic* features of discursive meaning groups words whereupon judges show high degrees of agreement with respect to relevance, but these classes of words group very heterogeneous words. In this section we try to combine semantic and formal features to find classes of words

- whereupon judges show good degrees of agreement,
- grouping words that are homogeneous with respect to how relevant they are judged by humans, and
- with a reasonable coverage of the corpus.

We have carried out a preliminary study about the combination of features, studying the behaviour with respect to ratio of reduction and inter-judge agreement for some combinations of semantic and formal features. In this preliminary study we have considered 17 classes, combining **punctuation** and **discourse markers** with,

- the features of the *semantic* dimension of discursive meaning (*context, equality, cause, revision*),
- hypotactic constructions (*participle, relative, apposition*)
- discourse markers and grammaticalized discourse markers

In Table 5.16 we can see the mean probability that a word is removed for each of these classes. Between parentheses we can see the probability that a word is removed in the corresponding class without the co-occurrence of punctuation and/or discourse markers. The same information is displayed graphically in Figures 5.8 and 5.9, the second one focuses the differences in combinations of semantic features with surface form features. We can see that, in virtually all the cases the probability that words are removed is bigger in classes characterized by more than one feature than in their corresponding classes with a single feature. For formal features the increase is rather small, but it is rather noticeable for semantic features, especially in their co-occurrence with punctuation.

Interestingly, only one exception can be found to this increase: words within the class of segments characterized by *revision* present higher probability of being removed when they do not co-occur with any other feature. We suspect that this is due to the fact that this is the most marked feature in the dimension of *semantic* discursive meaning, so additional marking cannot be accumulated, but has an interpretative effect. However, this claim is rather speculative and further research is needed to confirm it.

As seen in Table 5.17 and graphically in Figures 5.10 and 5.11, the kappa coefficient shows that the accumulation of evidences increases the agreement between judges, although no general explanation can be found for this increase. For example, co-occurrence with punctuation significantly increases agreement between judges for words within the classes of

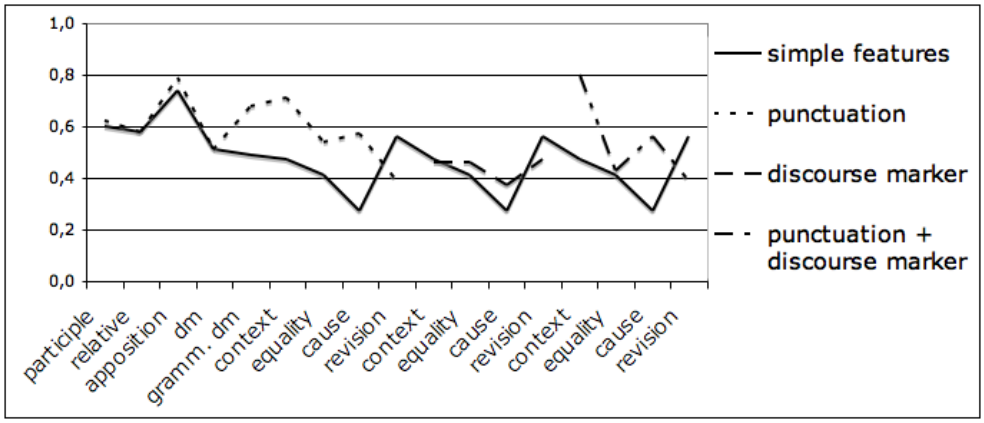


Figure 5.8: Differences in compression rates for segments characterized by a single feature vs. segments characterized by various features.

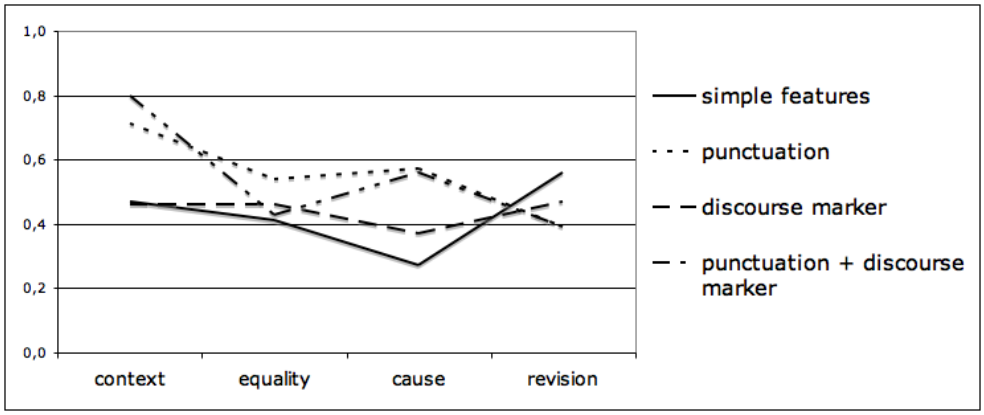


Figure 5.9: Differences in compression rates for segments characterized by semantic features alone or in combination with surface form features.

	average	a1	a2	a3	a4	a5
total	**.40**	.56	.44	.33	.38	.30
punctuation						
+ participle	**.62** (.60)	.77	.58	.50	.66	
+ relative	**.58** (.58)		.50	.63	.56	.66
+ apposition	**.79** (.74)	.97	.71	.67	.86	.78
+ disc. mark.	**.51** (.51)	.59	.54	.60	.44	.38
+ gram. disc. mark.	**.68** (.49)	.72	.55		.79	
+ context	**.71** (.47)	.84	.63		.73	.66
+ equality	**.54** (.41)	.89	.45	.45	.38	
+ cause	**.57** (.27)		.50		1	.23
+ revision	**.39** (.56)		.23	.50	.33	.52
discourse marker						
+ context	**.46** (.47)		.76	.42	.41	.28
+ equality	**.46** (.41)	.53	.47	.47	.40	
+ cause	**.37** (.27)		.46	.03	.73	.27
+ revision	**.47** (.56)		.37	.50	.49	.52
punctuation and discourse marker						
+ context	**.80** (.47)		.84		1	.57
+ equality	**.43** (.41)	.40	.46	.51	.36	
+ cause	**.56** (.27)		.47		1	.23
+ revision	**.39** (.56)		.23	.50	.33	.52

Table 5.16: Ratio of reduction of words occurring in different kinds of discourse segments distinguished by their semantics and their form. Between parenthesis, the ratio of reduction of each feature alone, without the co-occurrence of punctuation or discourse markers.

participle and *grammatical discourse marker*, but produces virtually no difference for other classes, probably because they usually co-occur with punctuation (*apposition*), because they are already highly marked inherently (*discourse markers*) or because their behaviour is due to other factors, like determination (*relative*).

As detailed in Figure 5.11, semantic features are affected differently by the co-occurrence with other evidence: the most marked case, *revision*, presents higher measures of agreement when co-occurring with punctuation, while it shows slightly less agreement when co-occurring with discourse markers. The least marked case, *context*, presents the opposite pattern: higer measures of kappa when co-occurring with discourse markers or with discourse markers and punctuation, and agreement drops if only co-occurrence with punctuation is considered. The same holds for *equality*, a less marked case as well, but with smaller differences in agreement. *Cause* shows higher measures of agreement when co-occurring with any kind of evidence.

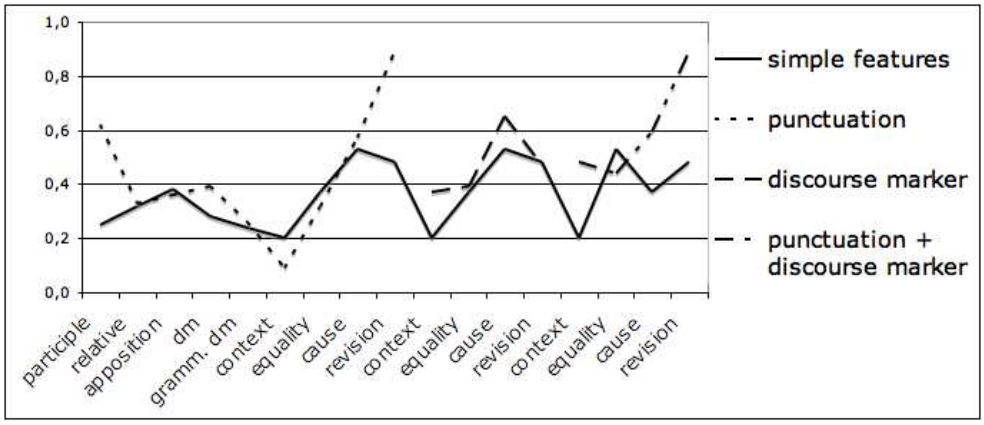


Figure 5.10: Differences in the kappa coefficient for inter-judge agreement for segments characterized by a single feature vs. segments characterized by various features.

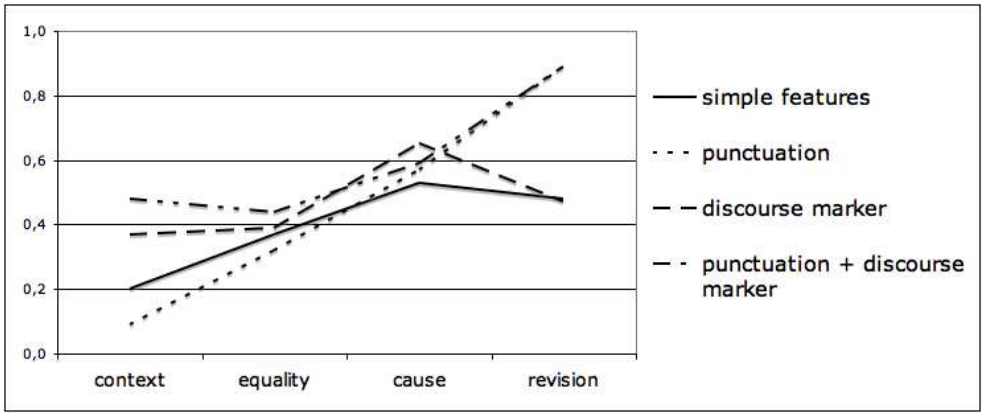


Figure 5.11: Differences in the kappa coefficient for inter-judge agreement for segments characterized by semantic features alone or in combination with surface form features.

	majority opinion	overall agreement	specific removability	specific non-removability	Kappa
total	.82	.70	.56	.76	.28
punctuation					
+ participle	.66	.45	.44	.15	**.62 (.25)**
+ relative	.70	.48	.50	.35	**.33 (.32)**
+ apposition	.82	.68	.76	.10	**.36 (.38)**
+ gram. disc. mark.	.83	.69	.77	.36	**.39 (.28)**
+ disc. mark.	.76	.62	.62	.54	**.26 (.24)**
+ context	.79	.62	.72	.29	**.09 (.20)**
+ equality	.83	.71	.68	.61	**.32 (.37)**
+ cause	.89	.80	.71	.49	**.57 (.53)**
+ revision	.70	.53	.27	.50	**.89 (.48)**
discourse marker					
+ context	.78	.64	.55	.58	**.37 (.20)**
+ equality	.81	.69	.67	.67	**.39 (.37)**
+ cause	.89	.82	.54	.84	**.65 (.53)**
+ revision	.69	.52	.43	.48	**.47 (.48)**
punctuation and discourse marker					
+ context	.87	.75	.82	.11	**.48 (.20)**
+ equality	.76	.62	.52	.60	**.44 (.53)**
+ cause	.89	.81	.71	.50	**.59 (.37)**
+ revision	.70	.53	.27	.50	**.89 (.48)**

Table 5.17: Inter-judge agreement measures for removability, distinguishing segments by their semantics. Between parenthesis, the kappa coefficient of each feature alone, without the co-occurrence of punctuation or discourse markers.

From this preliminary study combining different features to characterize classes of words we can conclude that combining different features may contribute to improve our model of the behaviour of human judges in removing words, because we have obtained classes where inter-judge agreement is rather high. However, this study can only be taken as a starting point. Two main enhancements seem crucial to obtain a model that is reliable enough to be used in real-world applications: create a bigger corpus, so that the representativity of word classes is bigger and we can evaluate whether they actually contain heterogeneous or homogeneous words, and describe the behaviour of all possible combinations of features.

From this study of a corpus of manually annotated texts, it can be concluded that modelling the discursive structure of texts allows us to better model the probability that a word is removed, and that the judgements of humans on the relevance of words seem to be a good source of information to infer and test such model.

Indeed, judges agree slightly more than what could be expected by chance, so that the null hypothesis that they agree by mere chance can be rejected. More interestingly, if we observe the behaviour of judges in removing words, we can find groups of words that tend to be removed, words that tend not to be removed and words that different judges consider differently. These groups of words seem to correspond to some of the classes we proposed in the theory of discourse organization presented in the previous chapter.

This seems to indicate that the task of eliminating intrasentential chunks from a text according to their relevance can benefit from our theory of the organization of discourse. Future work includes studying in depth the combination of different kinds of information we have considered (semantic and of form) to obtain a classification of word classes that maximizes the estimation of the probability that a word is removed and intra-class agreement, or that at least allows us to identify classes with high inter-judge agreement. In any case, working with a bigger corpus and a bigger number of judgements will definitely improve the statistical significance of the results, and it will allow to drive stronger and more accurate conclusions.

5.3 Human judgements on discourse relations

In this section we present the manual identification and description of discourse relations in a corpus of newspaper articles. The purpose of this annotation was twofold: first, we wanted to test the empirical adequacy of the proposed relations of discourse, seeing if naive judges could significantly agree on discourse relations. On the other hand we wanted to consolidate an annotation procedure to create a bigger annotated resource in the future.

In what follows we describe the corpus and judges, annnotation schema, relating it with previous work, and then we discuss the consistency of the resulting annotation.

5.3.1 Corpus and judges

Three independent judges identified relations between discourse segments after being instructed about the annotation schema described in the following section. Of the three judges, two were naive judges, with no background in linguistics, and a linguist. The purpose of having naive judges annotate the corpus was to check the intuitivity of the proposed features. After each text was annotated, the annotation was discussed between judges and helpful criteria were included in the annotation schema, but the annotated text was not modified.

The corpus consisted on six newspaper articles, with some variation in subgenres: television critic, everyday stories, and opinion. We chose a varied corpus in order to check the

applicability of features in different kinds of text. The corpus is annotated in XML (see Figure 5.12), and can be found at `http://lingua.fil.ub.es/~lalonso/discor/`.

Texts had been manually pre-segmented in discourse minimal units, discourse segments and discourse markers. Judges had to identify the relations they held with other segments and the meanings of these relations.

The corpus consists of 6 articles (one of which used for training) totalling 3541 words (from 207 to 1042 words) in 154 sentences. There are 468 intrasentential segments, of which 84 are required by the argumental structure of the verb (mainly subjects and objects of verbs of thinking or saying) and the rest are different kinds of adjuncts, of which 226 are dominated by a discourse marker. 261 discourse markers have been found, corresponding to 101 different forms, the most frequent are *y* (*and*, 50 occurrences), *para* (*for, to*, 20 occurrences), *como* (*like, as*, 15 occurrences) and *pero* (*but*, 9 occurrences). 56 were relative clauses, 43 prepositional phrases, 21 list members, 16 absolute participle constructions (past and present), and so forth.

5.3.2 Annotation schema

In contrast with the experiment presented in the previous section, the identification of relations between discourse segments requires some training, because the task is not a natural task of interpretation of language. So, the judges had to be instructed with respect to the kind of relations to be identified in text and the inventory of meanings to tag relations. Even if the base concepts are intuitive, their formalization has to be elicited in order to make annotations comparable. In order to carry out this training of the judges, we developed an annotation schema.

Various annotation schemata have been developed in order to improve the consistency of annotation and to reduce the annotation cost (Discourse Resource Initiative 1997; Cooper *et al.* 1999; Carlson and Marcu 2001). Existing schemas are mostly based in the Rhetorical Structure Theory (RST) (Mann and Thompson 1988), at different levels of granularity and for different languages (Carlson *et al.* 2003; Potsdam Corpus 2004) or are focussed in the annotation of dialogue (Carletta *et al.* 1996; Core and Allen 1997).

In contrast with other approaches, we did not carry out an intensive training of judges, but decisions were taken basically relying on their intuitions, in order to test the intuitivty of the proposed features. However, as an aid to guide decision-taking, an annotation manual was created (Alonso *et al.* 2004a), and a preparatory text was annotated collaboratively. The manual was enhanced and refined during annotation, with issues that were specially controversial, like the annotation of discontinuous discourse markers (*so... that...*) or an effective procedure to systematize the assignment of discursive features to syntactical constructions like relative clauses or absolute participles.

The procedure for annotation was as follows: once a judge had read the whole text, discourse segments were characterized one by one in order of occurrence in the text, by the following features: **node(s) of attachment**, **features of meaning** and **glosses** for the selected features.

[1.2[1.2.1[1.2.1.1 Pese a] sus apellidos,] Pedrosa Urquiza no debía de ser un "buen vasco"][...] [1.7 Quizás por eso le han matado] [...]

[1.2[1.2.1[1.2.1.1 *Despite*] *his surnames,*] *Pedrosa Urquiza mustn't be a "good basque"*] *[...]* [1.7 *Maybe that's why he has been killed*] *[...]*

```
<seg id="1.2"     type="sentence">
   <AttachmentPoint="1.7" sym=""      asym="1"
    cause="1"    context=""  equality=""
    revision=""  elabora="1" progre=""
    GlossCause="because"
    GlossElabora="he has been killed">

   <seg id="1.2.1"   type="segment">
     <AttachmentPoint="1.2" sym=""      asym="1"
      cause=""     context=""  equality=""
      revision="1" elabora="1"  progre=""
      GlossRevision="if you have such surnames,
                  then you are a good basque"
      GlossElabora="Urquiza couldn't be
                 a good basque">

     <seg id="1.2.1.1"       type="disc_mark">
       <AttachmentPoint="1.2,1.2.1" sym=""   asym="1"
        cause=""           context=""  equality=""
        revision="1"       elabora="1" progre=""
        GlossRevision="if you have such surnames,
                    then you are a good basque"
        GlossElabora="Urquiza couldn't be
                   a good basque">

        ____Despite___________
    </seg>

    ____his_________________
    ____surnames____________
 </seg>
 ...
</seg>
```

Figure 5.12: Example of XML annotation of a sentence (1.2), a sub-sentential discourse segment (1.2.1) and a discourse marker (1.2.1.1).

5.3.2.1 Node of attachment

The value for this attribute depends on the kind of discourse unit: segments are assigned the identification number(s) of the node(s) where they are attached in the tree-like structure of discourse. In contrast, discourse markers are assigned the identification number of the nodes they relate. These nodes can be minimal discourse segments or discourse units constituted by continuous spans of discourse segments.

However, segments with a relating function (as in example (12)) are assigned the same kind of value as a discourse marker, that is, the identification numbers of the segments they relate. Contrastively, discourse markers that do not relate discourse segments, like those relating the text with the author (for example *luckily* or *of course*), can be attached to the discourse segment where they are found.

(12) No satisfecho con todo ello, el ciudadano vasco Urquiza desoía a su partido y cada día se tomaba sus potes y vermuts en el "batzoki" del PNV de Durango.
Not happy with all this, the basque citizen Urquiza didn't listen to his party and he had his "potes" and vermouths in the "batzoki" of the PNV in Durango.

Each segment is related to at least one other segment. When there is no clear attachment node or when the segment is the starting point or main claim of a text, it is considered the top of a local structure, and is attached to itself.

Each segment carries the information for its own attachment to other segments, but none for the attachment of other segments to it. The most marked segment in a relation is the one that carries the information, usually, segments under the scope of a discourse marker or characterized by redundant lexic or referential expressions. If there is no difference in markedness between segments, those coming later in discourse are the ones to carry the information. Note that markedness has no relation with RST's nuclearity, since marked segments can be the nucleus or the satellite of a relation.

5.3.2.2 Features of meaning

We established eight basic features to describe the meaning of discourse relations, summarized in Table 5.18. One of the aims of this corpus annotation was to test the preliminary inventory of discursive meanings in order to settle the final inventory presented in Chapter 4. Two of the features of the preliminary inventory (*symmetric* and *asymmetric*) were found irrelevant after this preliminary experiment and were not included in the final inventory. Features were organized in three dimensions of meaning inferred from co-occurrences of prototypical discourse markers (Alonso *et al.* 2003d):

sentence-structural (*symmetric, asymmetric*) percolation of sentential syntax to discourse, absent from our final proposal for reasons of economy and adequacy of representation discussed in Section 4.2.2.

discourse-structural (*continuation, elaboration*)

semantic (*context, causality, equality, revision*)

feature	discursive effect
context	provides the setting for a discourse entity *People started demonstrating as soon as the war began.*
equality	establishes an equivalence between two elements *A is (mod) B* *Some other governments supported the war, as in Spain.*
causality	elicits a causal relation between two elements *They lost the elections because they manipulated information.*
revision	negates some previous information, explicit or implied *No weapons of mass destruction were found, but Iraq was invaded.*
continuation	introduces a new topic or intention *The "Prestige" wandered about for a week, and it finally sunk.*
elaboration	continues a presented topic or intention *The "Prestige" wandered about for a week, all along the coast.*
symmetric	attachement to a node at the same level in the discourse tree *They lied to voters and so they lost the elections.*
asymmetric	attachment to a node in a different level in the discourse tree *Because they lied to voters, they lost the elections.*

Table 5.18: Basic components of meaning of discourse relations, in three dimensions.

Within each of these dimensions, the range of possible meanings is ordered in a scale of markedness, from a default to the most marked case. The procedure to determine which meaning holds in each of this dimensions is systematized via the decision trees presented in Section 4.2.3. When judges were unsure about a feature for a relation, they left it underspecified. This resulted in an average of 2.6 features per relation, which increases slightly (2.7) for segments containing a discourse marker and decreases to 2.2 for relations at the beginning of a paragraph.

5.3.2.3 Glosses

If a *semantic* or *discourse-structural* feature was applicable for a relation, a gloss was provided for it. Glosses were aimed to guarantee consistency and keep track of decision procedures, but they also served to encode finer-grained distinctions within each of the proposed features, which will be exploited to study the realization of discursive meanings that are widely accepted in the literature in a bigger corpus. The following glosses were provided (examples can be seen in Figure 5.12):

revision elicit the information that is denied; in order of markedness, the first three based in Lagerwerf (1998), the least marked based in Umbach (2004):

denial of expectation elicit expectation
concessive elicit *tertium comparationis*
opposition the related segments can be rephrased with a correction *but* (*sino*)
focus-based if none of the others apply

cause one of four discourse markers with which the relation can be paraphrased: *in order to* (purpose), *because* (cause), *that's why* (reason), and *therefore* (consequence)

equality the common class of things to which the two segments belong, which has to be salient in the context, either because it is lexicalized or by regular abstraction procedures, like hyperonymy

Glosses for **continuation** and **elaboration** summarize the topic or intention that is introduced or elaborated, respectively. This procedure increased the consistency of the annotation for these two features

5.3.3 Preliminary results of annotation

First of all, we studied how each possible pair of features co-occurred, to detect possible redundancies in the features to describe the semantics of discourse relations. We provided the correlation coefficient for each pair of features, ranging from 1 to -1, where 1 indicates that the two features of a pair always co-occur in the same nodes, and -1 indicates that they never co-occur. As can be seen in Table 5.19, the correlation between features belonging to the sentence- and discourse-structural dimensions present a very high correlation: symmetric tends to co-occur with continuation and elaboration with asymmetric, and almost never the reverse. This, together with the fact that systematic mappings can be found between sentence- and discourse-structural features, seems to indicate that sentence-structural properties are irrelevant for the description of the semantics of discourse relations.

The consistency of the annotation was evaluated by kappa agreement between the values of the components of meaning assigned to each node. Agreement for point of attachment was very low and therefore it is not significative to provide it here. Additionally, we calculated the agreement for those nodes corresponding to segments dominated by a discourse marker and to segments at the beginning of paragraph.

As can be seen in Table 5.20, the average agreement, $\kappa = .54$ is quite low, and does not guarantee a good reproducibility of the results. However, it has to be taken into account that this agreement has been obtained in the preliminary phase of annotation, during the process when annotation criteria were being established, and also that the judges were not professional annotators. Even with trained professional annotators, Carlson, Marcu and Okurowski (2003) present $\kappa = .6$ in the initial stages of annotation, reaching $\kappa = .75$ at the end of the project.

Interestingly, kappa agreement for segments dominated by a discourse marker is significantly lower than the average, except in the case of the *revision* feature, which presents the highest agreement per feature. However, the number of features for characterizing discourse markers was higher than the average, which means that judges recognized them as highly informative of discourse organization but perceived the components of their meaning differently.

As can be seen by the low agreement for segments occurring at the beginning of paragraph, they were very difficult to characterize for judges, which seems to indicate that the

	causality	**equality**	**revision**	**elab**	**cont**	**symm**	**asymm**
context	-.15	**-.47**	-.04	.47	-.43	-.22	.27
causality	-	-.21	-.06	-.14	.15	-.15	.17
equality	-	-	-.05	-.39	.44	.52	-.47
revision	-	-	-	-.09	.06	.07	-.09
elaboration	-	-	-	-	**-.92**	-.52	.56
continuation	-	-	-	-	-	.57	-.52
symmetric	-	-	-	-	-	-	**-.94**

	causality	**equality**	**revision**	**elab**	**cont**	**symm**	**asymm**
context	.13	-.15	-.07	.17	-.19	-.16	.18
causality	-	-.19	.07	-.14	.16	.08	-.01
equality	-	-	-.16	.10	-.02	.25	-.19
revision	-	-	-	-.26	.32	.32	-.29
elaboration	-	-	-	-	**-.85**	-.49	.54
continuation	-	-	-	-	-	.54	-.46
symmetric	-	-	-	-	-	-	**-.89**

Table 5.19: Confusion matrices of the correlation coefficient between the distribution of features of meaning in a descriptive (above) and in an argumentative (below) texts (average kappa agreement between annotators $kappa = .64$ and $kappa = .54$, respectively).

coherence mechanisms that apply at paragraph level are qualitatively different from those at sentential and inter-sentential level. This seems a strong argument to treat the low level organization of discourse as an autonomous level of language.

The fact that *revision* presents the highest agreement of all features seems to provide empirical support to the hierarchy of markedness proposed in Section 4.2.3.3: since revision is the most marked case, it is perceived more clearly by all judges, which results in higher agreement with respect to this meaning. This would also explain the low value of agreement for *context*, the lowest of all cases: since it is the least marked case, it is only weaakly perceived by judges, which makes them disagree. However, it can also be thought that the low value of agreement for context is due to the fact that context is not a well-defined area of meaning, as discussed in Section 4.2.3.3.4.

Although judges had been allowed to relate segments to more than one discourse unit by more than one relation, this never happened in the corpus. This supports the claim that discourse can be represented as a hierarchical tree. In our framework, the structure of discourse is described as a superposition of the hierarchical trees that represent each dimension, where segments have only one relation in each dimension, although they might be attached to different segments in different dimensions (see Figure 3.1).

revision	cause	equality	context	prog.	elab.	sym.	asym.	**average**
0.70	**0.55**	**0.57**	**0.43**	**0.51**	**0.48**	**0.55**	**0.54**	**0.54**

discourse markers								
0.72	**0.53**	**0.53**	**0.39**	**0.44**	**0.43**	**0.39**	**0.42**	**0.48**

segments starting paragraph								
0.03	**-0.0**	**0.16**	**0.01**	**0.29**	**0.18**	**0.03**	**0.11**	**0.10**

Table 5.20: Average kappa agreement between judges for the six annotated texts.

In sum, in this section we have presented the application of the proposed feature-based approach to the description of the organization of discourse to corpus annotation. An annotation framework has been established to systematize the procedures for decision taking, implemented as decision trees. These procedures are a direct mapping of the internal organization of the dimensions of discursive meaning proposed in Chapter 4.

The consistency of this preliminary annotation is comparable to that of the preliminary stages of other annotation initiatives. The fact that two of the three annotators involved had no background in linguistics provides support for the validity of the proposed features as basic components of the meaning of discourse relations.

A detailed analysis of the data allowed to detect redundancies in the initial set of features, which was then reduced to only those meanings that were found relevant for distinguishing discursive meanings. Moreover, the markedness hierarchy proposed for the semantic dimension of meaning is validated by the fact that more marked features present higher values of agreement.

This is only the preliminary stage of the development of a corpus annotated with discursive information. With a bigger corpus, we will be able to provide more significant descriptions for the phenomena under inspection, like a more thorough study of inter-judge agreement, an exploration of the patterns of attachment, the study of the glosses, to get a grasp of the patterns of agreement of subkinds of revision, cause, topics, etc.

5.4 Automatic identification of discourse segments

There have been many attempts to identify segment boundaries automatically (Hirschberg and Grosz ; Hirschberg and Litman 1993; Passonneau and Litman 1997a; Di Eugenio *et al.* 1997). In most cases, algorithms are induced from a manually annotated corpus, where segments and their characterizing features are encoded by human judges. A number of features, both linguistic and nonlinguistic, have been used to characterize discourse segments: prosody, cue phrases, referential links, intentional and informational status of segments, types of relations, level of embedding, etc.

In our case, we had only little data from human judgements, clearly insufficient to induce an algorithm by machine learning techniques. Therefore, we took the corpus segmented by human judges only as a source of evidence to build an algorithm for the automatic identification of discourse segments. The basic structure of the algorithm is presented in Section 5.4.1.

The analysis of the human segmentation experiment presented in the previous section seems to indicate that there is a strong relation between shallow evidence on the organization of discourse and human judgements (or part of them). This is very convenient to our shallow approach to discourse analysis, and provides support to the development of a segmentation algorithm that is fully based on shallow evidence, as the one we present here.

We present two versions of the algorithm: one that exploits evidence with no linguistic pre-processing and another that relies on a shallow linguistic analysis (chunking). We show that these two versions of the algorithm are perfectly compatible with each other, since they are organized incrementally. This incremental architecture also allows for future enhancements of the algorithm to be added naturally, for example, if full parsing or semantic parsing are to be incorporated.

We present a rough evaluation of this algorithm in Section 5.4.3, purely *indicative* of the performance of the algorithm and the reliability of different kinds of information. We have evaluated the algorithm by standard precision and recall measures by comparison to two gold standards based on the set of texts used for the manual annotation presented in the previous section:

- the corpus where we manually identified *marked* and *unmarked* segments according to the theoretical concept of discourse segment presented in Section 3.3.2.

- a corpus where marked segments correspond to those words removed by the majority of judges, described in Section 5.2.

5.4.1 Algorithm to identify discourse segments

Following the general definition of segments proposed in Section 3.3.2, the algorithm has two distinct aims: first, identifying discourse segment boundaries and assessing their reliability as such. Second, checking that the strings contained within boundaries satisfy the conditions on content necessary to be considered discourse segments. A formal outline of the algorithm can be seen in Figure 5.13; a description of the main steps follows.

Two basic actions can be distinguished in the algorithm: the default case, continuing a segment (*Push Word to Segment*), and the marked case, inserting a segment boundary (*Make Segment Boundary* and *Insert Segment*), which is only realized when there is evidence that indicates the presence of a boundary above a certain reliability threshold and the requirements for completeness of the segment are met, if any.

Make Segment Boundary at the position n (w_n) in the Input String: consider w_n (punctuation, word or a cluster of any of them) a segment boundary, so that the

```
for each Word in Text

    if BoundaryCandidate(Word)
        Explore Context
        if Reliability(Context) > Reliability Threshold
          and Complete(Segment)
            if Parenthetical(Word)
                Insert Segment
            else Make Segment Boundary
        else Push Word to Segment
    else Push Word to Segment
```

Figure 5.13: Algorithm for automatic identification of discourse segments.

previous segment is defined as $S = ..., w_{n-3}, w_{n-2}, w_{n-1}$ and the next segment is defined as $S = w_{n+1}, w_{n+2}, w_{n+3}, ...$ Whether w_n belongs to the segment to its left, to its right or to none of them depends on the kind of element that is marking the sentence boundary.

Insert Segment insert a segment S_p from the position of the first parenthetical evidence (parenthesis, apposition mark) to the position of the second parenthetical evidence, without interrupting the containing segment S_c, so that if the first parenthesis is found at w_n and the second parenthesis is found at w_o, $S_p = w_n, ..., w_o$ and $S_c = ..., w_{n-1}, w_{o+1}, ...$

Segments are defined as lists of words organized in a sequence. Only in highly marked cases can a segment be embedded in another segment, so that the containing segment is not finished but interrupted by the embedded segment. This is the case for segments marked by parenthetical punctuation, clearly marked appositions (following the pattern `noun phrase comma noun phrase comma`) and segments with phrasal scope, as in example (13). Relative clauses are not considered as embedded segments because it often happens that they are not embedded, as in example (14).

(13) It is also AI's practice to give its material to governments before publication for their views and additional information and the organization will publish these in its reports.

(14) Laughlin es un físico teórico que relaciona áreas tan dispares como la del plegamiento de las proteínas y la de la superconductividad de altas temperaturas.
Laughlin is a theoretical physicist who relates such disparate areas as proteine folding and superconductivity at high temperatures.

The algorithm consists of the following basic methods:

identify boundary candidates mainly punctuation and discourse markers (from an electronic lexicon, as that in Appendix A) are possible boundary candidates, but also some syntactic structures (relative clauses, some kinds of phrases) can be considered as such, if a syntactic analysis of the text is available.

assess the reliability of boundary candidates to be actual segment boundaries, that is, disambiguate boundary candidates with respect to their sentential or discursive function (such functions are exemplified in Section 4.1.1, examples (2-a) and (2-b), respectively). For each boundary candidate, we can determine how reliably it is performing a discursive function based in:

- intrinsic reliability of each boundary candidate, as associated to it by the algorithm (in the case of punctuation) or in a lexicon (in the case of discourse markers). In the lexicon presented in Appendix A, discourse markers are not intrinsically associated to different degrees of reliability, although this could arguably be an intrinsic property, for example, co-related with a grammaticalization index. Indeed, highly grammatical discourse markers (see Table A.7) are not included in the lexicon because they are highly ambiguous, and thus unreliable as signals of discursive function, but are exploited by the segmentation algorithm as boundaries of weak reliability.
- its context of occurrence in the text
  - co-occurrence with other boundary candidates, so that the more boundary candidates co-occur, the more reliably they are performing a discursive function.
  - properties of their context of occurrence, like position in the paragraph or sentence, or part of speech of the surrounding words or constituents, if that is available.

assess whether the segment is complete a segment is usually considered complete if there is evidence signalling that it can convey a proposition; typically, if it is headed by an independent verbal form (inflected (15) or non-inflected (16)). We can also exploit other cases that clearly convey a distinct unit of meaning about the state of affairs, like appositions (17) or dislocated constituents (18). As already discussed in Section 3.3.2, characterizing discourse segments by their content requires a certain analysis of the text. Therefore, in the version of the algorithm that is exclusively based on pattern-matching, no requirements on content can be exploited.

(15) In a breakthrough case on 20 March this year, four policemen in Guatemala City were convicted of the murder of a 13-year-old street child and sentenced to between 10 and 15 years' imprisonment.

(16) Others freed included William Masiku, detained since 1980, Brown Mmpinganjira, detained since 1986, Margaret Marango Banda, detained since 1988, and Blaise Machira, also detained since 1988.

(17) Goodluck Mhango, a veterinary surgeon arrested in September 1987, has been rejected for release by a committee established to review the cases of political detainees.

(18) At their trial, the judge is believed to have added 25 to each sentence specifically because the police had carried out the attack while operating in their official capacity.

The following thresholds of reliability are established to determine when a boundary candidate is signalling an effective boundary:

reliability 1 the lowest necessary to identify a discourse segment boundary, achieved by the presence of a single weak evidence of boundary (a weak punctuation sign, a highly ambiguous discourse marker); the evidence provided by boundaries with reliability 1 is only considered when recall is highly priorized over precision.

reliability 2 an accumulation of two weak evidences of boundary or a strong one.

reliability 3 an accumulation of three weak evidences of boundary or an accumulation of evidences at least one of which is strong; this configuration always marks a segment boundary. When precision is priorized over recall, only segment boundaries of reliability 3 are considered as effective boundaries.

Table 5.21 displays results for recall and precision in the identification of segments considering segment boundaries at reliability levels 1, 2 and 3. In general, it can be seen that the higher the reliability threshold, the higher the precision. But a higher reliability threshold reduces recall much more than it increases precision, as is reflected in the decreasing F-measure. Different reliability thresholds can be used for different purposes of analysis.

5.4.2 Variants of the segmentation algorithm

Two variants of the basic segmentation algorithm have been developed, based on different kinds of input: raw text or text analyzed with shallow parsing. These two variants are related so that the second is naturally incremented from the raw text version. This incremental architecture also allows for future enhancements of the algorithm to be added naturally, for example, if full parsing or semantic parsing are to be incorporated.

The basic variation between the pattern-matching algorithm and the general one is that the first does not perform any analysis of the content of discourse segments, and also that syntatctic structures are not considered as boundary candidates. An overview of the algorithm can be seen in Figure 5.14. The following textual cues are exploited:

weak evidence of discourse segment boundary weak punctuation (","), vague discourse markers, (listed in Table A.7), highly grammaticalized words like subordinating and coordinating conjunctions or relative pronouns.

strong evidence of discourse segment boundary discourse markers, strong punctuation (“.”, ”?”, ”!”, “:”, “;”), parenthetical punctuation.

The algorithm based in shallow parsing, seen in Figure 5.15 follows the basic structure of the general algorithm, but it can exploit richer information about the context of occurrence of boundary candidates. The analysis provides information about the part of speech of words and their organization in phrase-like units (chunks). This information allows to improve the process of assigning reliability and to determine the scope of segment boundaries with higher accuracy. The following cases can be identified by exploiting the shallow analysis of text:

- constituents (adverbial phrases and prepositional phrases) dislocated to the left constitute segments, as in example (18).
- appositions can be identified as the pattern `noun phrase comma noun phrase comma`, as in example (17).
- if a phrasal discourse marker (typically, a preposition) is not followed by a subordinating conjunction (like *that*), its scope is reduced to the prepositional phrase it syntactically dominates, and a segment is *inserted*, instead of making a segment boundary. In the pattern-matching algorithm we do not know which words in a sentence constitute phrases, so we cannot regulate the scope of discourse markers according to their syntactical type and they always have scope until the following segment boundary.
- unreliable boundary candidates, like commas and coordinating conjunctions, can reliably indicate boundaries in a context like `verb ... boundary candidate ... verb`, as in example (15).

The following example shows different analysis provided for a same sentence by each variant of the segmentation algorithm, the one based in pattern-matching (19-a) and the one based in shallow parsing (19-b), with a threshold of reliability $r = 2$. Boundary candidates are in boldface, candidates with $r > 1$ are boxed, and evidence provided by the context is italicized.

(19) AI delegations have been expelled from countries after discovering evidence of human rights abuse and several countries, China, have refused Amnesty entry.

a. AI delegations have been expelled from countries **after** discovering evidence of human rights abuse **and** several countries**,** China**,** have refused Amnesty entry.

b. AI delegations *have been expelled* from countries **after** discovering evidence of human rights abuse **and** *several countries, China, have refused* Amnesty entry.

```
for each Word in Text

    if Punctuation(Word)
        if Strong(Word)
            Reliability = 3
        else if Parenthetical(Word)
            Insert Segment
        else if Weak(Word)
             Reliability = 1
             if DiscourseMarker(Word+1)
                 Reliability = 3
             else if VagueDiscourseMarker(Word+1)
                 Reliability = 2

    else if DiscourseMarker(Word)
        Reliability = 2
            if DiscourseMarker(Word+1) or VagueDiscourseMarker(Word+1)
                Reliability = 3

    else if VagueDiscourseMarker(Word)
        Reliability = 1
            if DiscourseMarker(Word+1)
                Reliability = 3
            else if VagueDiscourseMarker(Word+1)
                Reliability = 2

    if Reliability(Context) > Reliability Threshold
        Make Segment Boundary
    else Push Word to Segment
```

Figure 5.14: Algorithm for automatic identification of discourse segments by pattern-matching.

```
for each Word in Text

    if Punctuation(Word)
        if Strong(Word)
            Reliability = 3
        else if Parenthetical(Word)
            Insert Segment
        else if Weak(Word)
             Reliability = 1
             if DiscourseMarker(Word+1)
                 Reliability = 3
             else if VagueDiscourseMarker(Word+1)
                 Reliability = 2
                 if HasVerb(Segment) and HasVerb(Segment+1)
                     Reliability = 3
             else if ( PrepPhrase(Chunk-1) or AdvPhrase(Chunk-1) )
               and Strong(Punctuation(Chunk-2))
                 Reliability = 3
             else if NounPhrase(Chunk-1) and NounPhrase(Chunk+1)
               and Punctuation(Chunk+2)
                 Insert Segment

    else if DiscourseMarker(Word)
        Reliability = 2
            if DiscourseMarker(Word+1) or VagueDiscourseMarker(Word+1)
              or AdvPhrase(Chunk-1) or PrepPhrase(Chunk+1)
                Reliability = 3
        if Phrasal(Word) and PrepPhrase(Chunk)
            and Reliability(Context) > Reliability Threshold
                Insert Segment

    else if VagueDiscourseMarker(Word)
        Reliability = 1
            if DiscourseMarker(Word+1)
                Reliability = 3
            else if VagueDiscourseMarker(Word+1)
              or AdvPhrase(Chunk-1) or PrepPhrase(Chunk+1)
                  Reliability = 2
            if HasVerb(Segment) and HasVerb(Segment+1)
                Reliability = 3

    if Reliability(Context) > Reliability Threshold
        Make Segment Boundary
    else Push Word to Segment
```

Figure 5.15: Algorithm for automatic identification of discourse segments in text analyzed

The natural enhancement to the algorithms presented here consists in exploiting information from full parsing. Unfortunately, full parsing is beyond the current state of the art NLP for Spanish and Catalan. In Fuentes *et al.* (2003) we applied this intuition for English, exploiting the full parse provided by MINIPAR (MINIPAR 1998), and showed that exploiting argumental structure is useful to compress sentences for headline style summaries.

We also showed that the relevance of words in the text was also indicative of the relevance of its containing segment, but is also indicative of whether a constituent can be considered a segment independent of its matrix clause or not, when no other information is available. In example (20), we do not know whether *in light of* is a discourse marker, because it is not stored in our lexicon, and we do not know either whether the present partciple *surrounding* is in absolutive position, because there is no comma preceding it. However, the fact that they contain words that are relevant in the document and different from those in the matrix clause (in boxes) allows to identify them as autonomous segments with high reliability.

(20) TORONTO (AP) Members of the delegation for [Quebec City]'s 2002 Winter Olympics [bid] feel betrayed in light of the [scandal] surrounding the successful [bid] by [Salt Lake City].

5.4.3 Evaluation of the automatic identification of discourse segments

In this section we present a rough evaluation of the variants of the algorithm for automatic identification of discourse segments. No strong conclusions can be obtained from this evaluation, for two main reasons: first, the evaluation corpus is rather small, so it has low statistical significance. On the other hand, as follows from the conclusions in Section 5.2, the concept of discourse segment is not stable empirically. Therefore, a gold standard based on this concept will also not be stable, and neither will be the evaluation based in a comparison with such gold standard.

Even though, this evaluation aims to be *indicative* of the performance of the general algorithm and its variants. Most of all, it aims to gain insight on the reliability of the different kinds of information that contribute to identify discourse segments.

We present an evaluation based in a comparison with two gold standards, measured by precision and recall, seen in Table 5.21.

5.4.3.1 Comparison with a gold standard

We have evaluated the algorithm by standard precision and recall measures with comparison to two gold standards based on the set of texts used for the manual annotation presented in Section 5.2:

– the corpus where we identified *marked* and *unmarked* segments according to the

theoretical concept of discourse segment(described in Section 5.2.1) (*theoretical* gold standard), and

– a corpus where segments were identified by taking the opinion of the majority of judges to determine which spans of text are removable (*removability* gold standard).

Comparisons between automatic and manual segmentation have been carried out in a per-word basis, instead of a per-segment basis. We have chosen this rate of comparison because it allows to compare segments produced by theoretical, removability and automatic criteria, which do not always cover the same spans of text.

Two different versions of each algorithm have been evaluated: one that is based on the lexicon of prototypical discourse markers presented in Section 3.4.3, and another one that is based on an *ad-hoc* lexicon of discourse markers that contains all the discourse markers present in the texts to be segmented. The aim of these two versions of the algorithms is to assess the contribution of discourse markers to identify discourse segments.

The results of evaluation can be seen in Table 5.21. Results are given in terms of precision, recall and their harmonic mean (F measure) by comparison to human segmentation. We can see that there is only a slight difference in the comparison between automatic procedures and any of the two gold standard, either theoretical or removability. This is good news, because it indicates that automatic procedures seem to identify segments that are perceived by all kinds of judges.

It can be seen that the algorithm based in pattern matching systematically presents more recall than the algorithm based in shallow parsing, but the latter obtains higher precision, although with a lower harmonic mean.

An extended discourse marker lexicon improves recall, but restrictions in reliability, that is, increasing the reliability index, produce a drop in recall, without a compensation in precision, which severly affects the harmonic mean. Recall is not improved by an extended discourse marker lexicon. The best results seem to be obtained at levels of precision of 1 or 2, and with an extended discourse marker lexicon.

It has often been shown (Litman 1996) that the ambiguity of textual cues can be solved if various sources of evidence are combined. We have successfully combined discourse markers, shallow syntactic structures and punctuation, which yields an improvement in the task of automatic discourse segmentation. We have also shown that discourse markers are the most informative of these three sources of evidence is discourse markers, and that their combination with punctuation significantly increases the reliability of the segmentation. Partial syntactical structures (chunks) provide information that is only reliable if combined with punctuation, and that is more error prone than discourse markers.

Hirschberg and Litman (1993) shows that the role of punctuation is equivalent to that of prosody to determine the function of discourse markers. A comparison of the textual and prosodic models proposed by Hirschberg and Litman show that the textual model, that exploits only punctuation, performs significantly better than the one relying on prosody (20% vs. 25% error rate, respectively). Therefore, there are both applied and theoretical reasons to consider punctuation as a first-class source of information about discourse structure.

algorithm based in pattern-matching

confidence threshold	**gold standard**					
	theoretical			**removability**		
	precision	recall	**F-measure**	precision	recall	**F-measure**
default discourse markers						
0	.69	.71	**.70**	.49	.67	**.57**
1	.73	.42	**.53**	.54	.42	**.47**
2	.70	.31	**.43**	.54	.32	**.40**
3	.78	.28	**.41**	.58	.28	**.37**
ad-hoc discourse markers						
0	.67	.78	**.72**	.47	.73	**.57**
1	.68	.53	**.60**	.49	.51	**.50**
2	.72	.34	**.46**	.56	.35	**.43**
3	.78	.29	**.42**	.54	.29	**.38**

algorithm based in shallow parsing

confidence threshold	**gold standard**					
	theoretical			**removability**		
	precision	recall	**F-measure**	precision	recall	**F-measure**
default discourse markers						
0	.78	.52	**.61**	.57	.51	**.52**
1	.78	.52	**.61**	.57	.51	**.52**
2	.78	.57	**.47**	.57	.46	**.50**
3	.87	.20	**.32**	.68	.22	**.32**
ad-hoc discourse markers						
0	.75	.60	**.66**	.54	.58	**.55**
1	.75	.60	**.66**	.54	.58	**.55**
2	.78	.52	**.62**	.58	.52	**.54**
3	.87	.22	**.35**	.66	.23	**.34**

Table 5.21: Evaluation of two algorithms, one based in patterh-matching and another based in shallow parsing for automatic discursive segmentation of text, by comparison with a gold standard based on the *theoretical* concept of discourse segment and by comparison with a gold standard based on the judgements about *removability* of words made by naive judges.

From all this we can conclude that automatic discouse segmentation at the level of minimal discourse units reaches a satisfactory performance level by shallow NLP techniques only. It benefits strongly from a good treatment of discourse markers, specially of their disambiguation with respect to sentential or discursive function and with respect to their relation with punctuation markers. This suggests that a work like that of Marcu (2000) to describe discourse markers is basic to discourse parsing. An extensive study of the contribution of each kind of evidence to the improvement of the automatic segmentation algorithm is left for future work.

Deeper linguistic analyses are of arguable use to improve discourse segmentation. Partial syntactical structures do not seem to be useful for this kind of discourse segmentation. We have also experimented with using full syntactic structures to identify discourse segments exploiting argumental structure (Fuentes, Massot, Rodríguez and Alonso 2003), but the contribution of this information still remains to be analyzed in depth. Our intuitions indicate that other kinds of analyses, like information structure, would be of more use.

5.5 Automatic identification of discourse markers

As has been explained in the previous sections, discourse markers constitute a major source of evidence to improve the performance of basic algorithms to obtain automatically a representation of discourse, that is, to identify discourse segments and the relations between them. Therefore, applications will work better if they have a bigger amount of discourse markers.

Work concerning discourse markers has been mainly theoretical, and applications to NLP have been mainly oriented to restricted natural language generation applications, mainly for English and, to some extent, also for German. The usual approach to building discourse marker resources is fully manual. For example, discourse marker lexicons are built by gathering and describing discourse markers from corpus or literature on the subject, a very costly and time-consuming process. Moreover, due to variability among humans, discourse marker lexicons tend to suffer from inconsistency in their extension and intension. To inherent human variability, one must add the general lack of consensus about the appropriate characterisation of discourse markers for NLP. All this prevents reusability of these costly resources.

As a result of the fact that discourse marker resources are built manually, they present uneven coverage of the actual discourse markers in corpus. More concretely, when working on previously unseen text, it is quite probable that it contains discourse markers that are not in a manually built discourse marker lexicon. This is a general shortcoming of all knowledge that has to be obtained from corpus, but it becomes more critical with discourse markers, since they are very sparse in comparison to other kinds of corpus-derived knowledge, such as terminology. As follows, due to the limitations of humans, a lexicon built by mere manual corpus observation will cover a very small number of all possible discourse markers.

In Section 3.4 we have defined discourse markers, identifying their characterizing fea-

tures and creating a lexicon of prototypical discourse markers. In this section we will apply a knowledge-poor lexical acquisition approach to find previously unseen discourse markers from huge amounts text, with an acceptable precision rate. To do that, we will exploit some features of discourse that are not strictly defining, in the sense that they may not necessarily be applicable to all discourse markers and that they may be applicable to linguistic items that we do not consider discourse markers, but that we have found highly indicative of *markerhood*[3], thus contributing to complete the definition of *discourse marker*.

This section is organized as follows. First, we describe the characterizing features of discourse markers that we have exploited. Then we present X-Tractor, a tool that applies a knowledge-poor approach to identify discourse markers as characterized from their occurrences in huge amounts of text in Section 5.5.2. Section 5.5.3 discusses the results obtained by X-Tractor in an experiment to enhance a discourse marker lexicon for Spanish.

5.5.1 Features characterizing prototypical discourse markers

The features that have exposed in Section 3.3 allow a human judge to determine whether a given string is a discourse marker, but they are not helpful to discriminate what might be a discourse marker among a set of strings with shallow NLP techniques, because they rely on a deep understanding of text.

Discourse markers have a tendency to co-occur with certain phenomena at surface level, which can then be taken as indicators of the presence of a discourse marker. The phenomena that we have found indicative of markerhood are:

- occurrence at the beginning of paragraph, beginning of the sentence, surrounded by punctuation; these are structural positions that are preferrably occupied by words relating discourse segments.
- containing a previously known discourse marker or part of it; this signals that (at least part of) the semantics of the string is discursive.
- containing one of a set of pre-defined patterns of words (of the kind *preposition + pronoun*, *adverb + subordinating conjunction*, *coordinating conjunction + adverb*); which indicate that they are relating elements beyond propositional scope.
- presence of anaphoric expressions; which signal a relation beyond propositional scope.
- high mutual information of the words that constitute the discourse marker; which is indicative of its level of grammaticalization.
- presence of unfrequent words; which signal a discontinuity in text and may be correlated with a segment boundary.

[3]By analogy with *termhood*(Kageura and Umino 1996), as used in terminology extraction to indicate the likelihood that a term candidate is an actual term, we have called *markerhood* the likelihood that a discourse marker candidate is an actual discourse marker.

We consider that, although these features are not exclusively defining discourse markers, and discourse markers may occur without any of these features, but, when huge amounts of corpora are taken into consideration, they are strongly indicative of markerhood. We support this claim empirically with an experiment to enhance a starting lexicon of discourse markers.

5.5.2 A knowledge-poor approach to acquire discourse markers from corpus: X-TRACTOR

One of the main aims of this system is to be useful for a variety of languages, including those for which NLP resources are not available. Therefore, we have tried to remain independent of any hand-crafted resources, including annotated texts or NLP tools. Following the line of (Engehard and Pantera 1994), syntactical information is worked by way of patterns of function words, which are finite and therefore listable. This makes the cost of the system quite low both in terms of processing and human resources.

Focusing on adaptability, the architecture of X-TRACTOR is highly modular. As can be seen in Figure 5.16, it is based in a language-independent kernel implemented in perl and a number of modules that provide linguistic knowledge.

The input to the system is a starting discourse marker lexicon and a corpus with no linguistic annotation. discourse marker candidates are extracted from corpus by applying linguistic knowledge to it. Two kinds of knowledge can be distinguished: general knowledge from the language and that obtained from a starting discourse marker lexicon.

The discourse marker extraction kernel works in two phases: first, a list of all might-be-discourse markers in the corpus is obtained, with some characterising features associated to it. A second step consists in ranking discourse marker candidates by their likelihood to be actual markers, or *markerhood*. This ranked list is validated by a human expert, and actual discourse markers are introduced in the discourse marker lexicon. This enhanced lexicon can be then re-used as input for the system.

In what follows we describe the different parts of X-TRACTOR in detail.

5.5.2.1 Linguistic Knowledge

Two kinds of linguistic knowledge are distinguished: general and lexicon-specific. General knowledge is stored in two modules. One of them accounts for the distribution of discourse markers in naturally occurring text in the form of rules. It is rather language-independant, since it exploits general discursive properties such as the occurrence in discursively salient contexts, like beginning of paragraph or sentence. The second module is a list of stopwords or function words of the language in use.

Lexicon-specific knowledge is obtained from the starting discourse marker lexicon. It also consists of two modules: one containing classes of words that constitute discourse markers and another with the rules for legally combining these classes of words. In the application of this system to Spanish, we started with a Spanish discourse marker lexicon

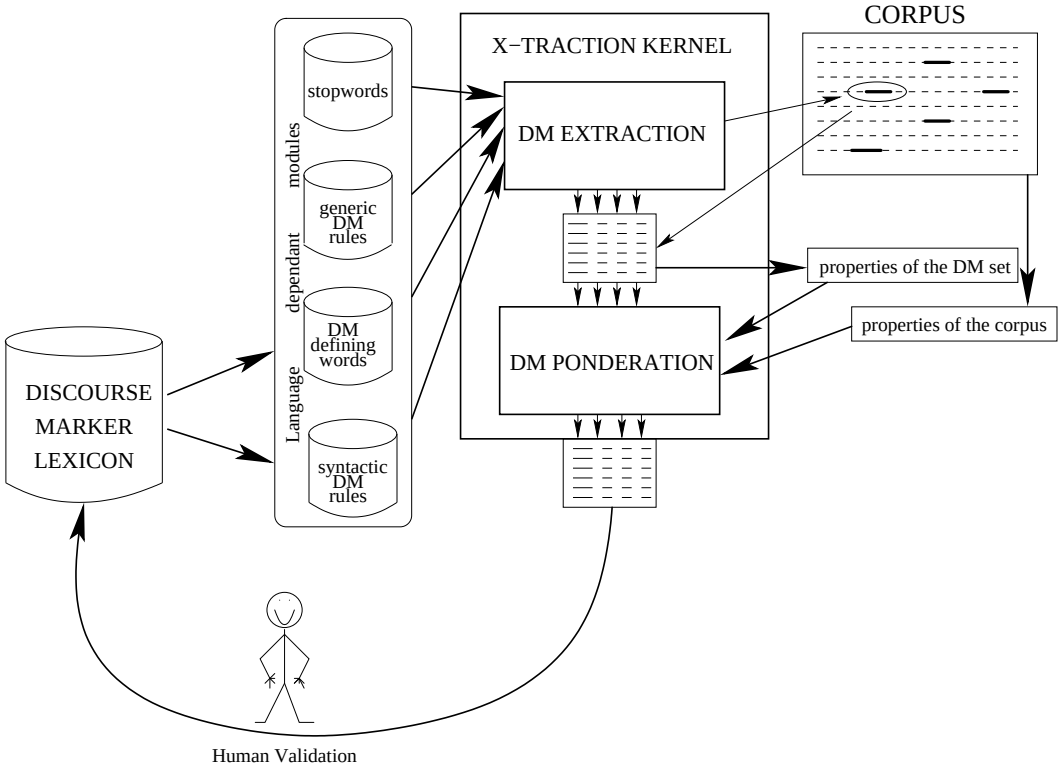


Figure 5.16: Architecture of X-Tractor

```
for each word in string
      if word is a preposition, then
        if word-1 is an adverb, then
          if word-2 is a coordinating conjunction, then
            if word+1 is a discourse-content word, then
              if word+2 is a preposition, then
                assign the discourse marker candidate structural weight 5
              elsif word+2 is a subordinating conjunction, then
                assign the discourse marker candidate structural weight 5
              else assign the discourse marker candidate structural weight 4
            elsif word+1 is a pronoun, then
              assign the discourse marker candidate structural weight 4
          else assign the discourse marker candidate structural weight 3
```

Figure 5.17: Example of rules for combination of discourse marker-constituing words

consisting of 577 discourse markers [4], described in Alonso (2001). This experiment was carried out before the lexicon described in Appendix A existed, but all discourse markers contained in our current prototypical lexicon were already present in this first lexicon.

We transformed this lexicon to the kind of knowledge required by X-Tractor, and obtained 6 classes of words (adverbs, prepositions, coordinating conjunctions, subordinating conjunctions, pronouns and content words), totalling 603 lexical items, and 102 rules for combining them. For implementation, the words are listed and they are treated by pattern-matching, and the rules are expressed in the form of *if - then - else* conditions on this pattern-matching (see Table 5.17).

5.5.2.2 Extraction of discourse marker candidates

Discourse marker candidates are extracted by applying the above mentioned linguistic knowledge to plain text. Since discourse markers suffer from data sparseness, it is necessary to work with a huge corpus to obtain a relatively good characterisation of discourse markers. In the application to Spanish, strings were extracted if they accomplished at least one of the following conditions:

- significant tendency to occur in positions typically occupied by words relating discourse segments: beginning of paragraph, beginning of the sentence, marked by punctuation.
- it contains lexical items that are parts of discourse markers in the lexicon.
- it matches one of a set of pre-defined patterns of words that signal that the string has a relating function beyond propositional scope.

[4] We worked with 784 expanded forms corresponding to 577 basic cue phrases

5.5.2.3 Assessment of markerood

Once all the possible might-be-discourse markers are obtained from corpus, they are ponderated as to their *markerhood*, and a ranked list is built.

Different kinds of information are taken into account to assess *markerhood*:

- **Frequency** of occurrence of the discourse marker candidate in corpus, normalised by its length in words and exclusive of stopwords. Normalisation is achieved by the function $normalised\ frequency = length \cdot \log(frequency)$.
- Frequency of occurrence in **discursively salient context**. Discursively salient contexts are preferred occurrence locations for discourse markers.
- **Mutual Information** of the words forming the discourse marker candidate. Word strings with higher mutual information are supposed to be more plausible lexical units.
- **Internal Structure** of the discourse marker, that is to say, whether it follows one of the rules of combination of discourse marker-words. For this application, X-Tractor was aimed at obtaining discourse markers other than those already in the starting lexicon, therefore, longer well-structured discourse marker candidates were priorised, that is to say, the longer the rule that a discourse marker candidate satisfies, the higher the value of this parameter.
- **Discourse Content** of the discourse marker candidate is increased by the number of words it contains that have clearly discursive semantics. These words are listed in one of the modules of external knowledge.
- **Lexical Weight** accounts for the the presence of non frequent words in the discourse marker candidate. Unfrequent words make a discourse marker with high *markerhood* more likely as a segment boundary marker.
- **Linking Function** of the discourse marker candidate accounts for its power to link spans of text, mostly by reference.
- **Length** of the discourse marker candidate is relevant for obtaining new discourse markers if we take into consideration the fact that discourse markers tend to aggregate.

These parameters are combined by weighted voting for *markerhood* assessment, so that the importance of each of them for the final *markerhood* assessment can be adapted to different targets. By assigning a different weight to each one of these parameters, the system can be used for extracting discourse markers useful for heterogeneous tasks, for example, automated summarisation, anaphora resolution, information extraction, etc.

In the application to Spanish, we were looking for discourse markers that signal discourse structure useful for automated text summarisation, that is to say, mostly indicators of relevance and coherence relations.

5.5.3 Results and discussion

We ran X-TRACTOR on a sample totalling 350,000 words of Spanish newspaper corpus, and obtained a ranked list of discourse markers with an indication of their potential as segment boundary markers. Only 372 out of the 577 discourse markers in the discourse marker lexicon could be found in this sample, which indicates that a bigger corpus would provide a better picture of discourse markers in the language, as will be developed below.

5.5.3.1 Evaluation of Results

Evaluation of lexical acquisition systems is a problem still to be solved. Typically, the metrics used are standard IR metrics, namely, *precision* and *recall* of the terms retrieved by an extraction tool evaluated against a document or collection of documents where terms have been identified by human experts (Vivaldi 2001). Precision accounts for the number of term candidates extracted by the system which have been identified as terms in the corpus, while recall states how many terms in the corpus have been correctly extracted.

This kind of evaluation presents two main problems: first, the bottleneck of hand-tagged data, because a large-scale evaluation implies a costly effort and a long time for manually tagging the evaluation corpus. Secondly, since terms are not well-defined, there is a significant variability between judges, which makes it difficult to evaluate against a sound golden standard.

For the evaluation of discourse marker extraction, these two problems become almost unsolvable. In the first place, discourse marker density in corpus is far lower than term density, which implies that judges should read a huge amount of corpus to identify a number of discourse markers significant for evaluation. In practical terms, this is almost unaffordable. Moreover, X-TRACTOR's performance is optimised for dealing with huge amounts of corpus. On the other hand, the lack of a reference concept for discourse marker makes inter-judge variability for discourse marker identification even higher than for term identification.

Given these difficulties, we have carried out an alternative evaluation of the presented application of the system. To give a hint of the recall of the obtained discourse marker candidate list, we have found how many of the discourse markers in the discourse marker lexicon were extracted by X-TRACTOR, and how many of the discourse marker candidates extracted were discourse markers in the lexicon[5]. To evaluate the goodness of *markerhood* assessment, we have found the ratio of discourse markers in the lexicon that could be found among the first 100 and 1000 highest ranked discourse marker candidates given by X-TRACTOR. To evaluate the enhancement of the initial set of discourse markers that was achieved, the 100 highest ranked discourse markers were manually revised, and we obtained the ratio of actual discourse markers or strings containing discourse markers that were not in the discourse marker lexicon. Noise has been calculated as the ratio of non-discourse

[5]We previously checked how many of the discourse markers in the lexicon could actually be found in corpus, and found that only 386 of them occurred in the 350,000 word sample; this is the upper bound of in-lexicon discourse marker extraction.

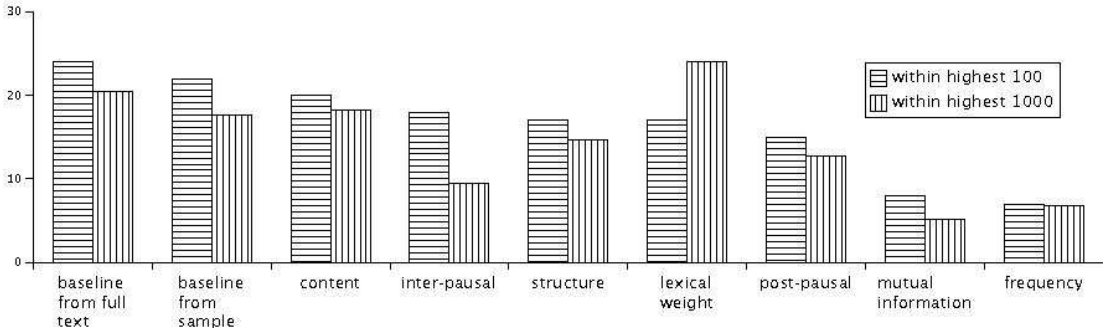


Figure 5.18: Ratio of discourse marker andidates that contain a discourse marker in the lexicon among the 100 and 1000 highest ranked by each individual parameter.

markers that can be found among the 100 highest ranked discourse marker candidates.

5.5.3.2 Parameter Tuning

To roughly determine which were the parameters more useful for finding the kind of discourse markers targeted in the presented application, we evaluated the goodness of each single parameter by obtaining the ratio of discourse markers in the lexicon that could be found within the 100 and 1000 discourse marker candidates ranked highest by that parameter.

In Figure 5.18 it can be seen that the parameters with best behaviours in isolation are *content*, *structure*, *lexical weight* and *occurrence in pausal context*, although none of them performs above a dummy baseline fed with the same corpus sample. This baseline extracted 1- to 4-word strings after punctuation signs, and ranked them according to their frequency, so that the most frequent were ranked highest. Frequencies of strings were normalised by length, so that $normalised\ frequency = length \cdot \log(frequency)$. Moreover, the frequency of strings containing stopwords was reduced.

	baseline	X-Tractor
Coverage of the discourse marker lexicon	88%	87.5%
ratio of discourse markers in the lexicon		
within 100 highest ranked	31%	41%
within 1000 highest ranked	21%	21.6%
Noise		
within the 100 highest ranked	57%	32%
Enhancement Ratio		
within the 100 highest ranked	9%	15%

Table 5.22: Results obtained by X-Tractor and the baseline

However, the same dummy baseline performed better when fed with the whole of the newspaper corpus, consisting of 3,5 million words. This, and the bad performance of the

parameters that are more dependant on corpus size, like *frequency* and *mutual information*, clearly indicates that the performance of X-Tractor, at least for this particular task, will tend to improve when dealing with huge amounts of corpus. This is probably due to the data sparseness that affects discourse markers.

This evaluation provided a rough intuition of the goodness of each of the parameters, but it failed to capture interactions beteween them. To assess that, we evaluated combinations of parameters by comparing them with the lexicon. We finally came to the conclusion that, for this task, the most useful parameter combination consisted in assigning a very high weight to structural and discourse-contextual information, and a relatively important weight to content and lengh, while no weight at all was assigned to frequency or mutual information. This combination of parameters also provides an empirical approach to the delimitation of the concept of discourse marker, by eliciting the most influential among a set of discourse marker-characterising features.

However, the evaluation of parameters failed to capture the number of discourse markers non present in the lexicon retrieved by each parameter or combination of parameters. To do that, the highest ranked discourse marker candidates of each of the lists obtained for each parameter or parameter combination should have been revised manually. That's why only the best combinations of parameters were evaluated as to the enhancement of the lexicon they provided.

5.5.3.3 Results with combined parameters

In Table 5.22 the results of the evaluation of X-Tractor and the mentioned baseline are presented. From the sample of 350,000 words, the baseline obtained a list of 60,155 discourse marker candidates, while X-Tractor proposed 269,824. Obviously, not all of these were actual discourse markers, but both systems present an 88% coverage of the discourse markers in the lexicon that are present in this corpus sample, which were 372.

Concerning goodness of discourse marker assessment, it can be seen that 43% of the 100 discourse marker candidates ranked highest by the baseline were or contained actual discourse markers, while X-Tractor achieved a 68%. Out of these, the baseline succeeded in identifying a 9% of discourse markers that were not in the lexicon, while X-Tractor identified a 15%. Moreover, X-Tractor identified an 8% of temporal expressions. The fact that they are identified by the same features characterising discourse markers indicates that they are very likely to be treated in the same way, in spite of heterogeneous discursive content.

In general terms, it can be said that, for this task, X-Tractor outperformed the baseline, suceeded in enlarging an initial discourse marker lexicon and obtained quality results and low noise. It seems clear, however, that the dummy baseline is useful for locating discourse markers in text, although it provides a limited number of them.

By this application of X-Tractor to a discourse marker extraction task for Spanish, we have shown that bootstrap-based lexical acquisition is a valid method for enhancing a lexicon of discourse markers, thus improving the limited coverage of the starting resource. The resulting lexicon exploits the properties of the input corpus, so it is highly portable

to restricted domains. This high portability can be understood as an equivalent of domain independence.

The use of this empirical methodology circumvents the bias of human judges, and elicits the contribution of a number of parameters to the identification and characterization of discourse markers. Therefore, it can be considered as a data-driven delimitation of the concept of discourse marker.

Future improvements of this tool include applying techinques for interpolation of variables, so that the tuning of the parameters for *markerhood* assessment can be carried out automatically. Also the process of rule induction from the lexicon to the rule module can be automatised, given classes of discourse marker-constituting-words and classes of discourse markers. Moreover, it has to be evaluated in bigger corpora.

Another line of work consists in exploiting other kinds of knowledge for discourse marker extraction and ponderation. For example, annotated corpora could be used as input, tagged with morphological, syntactical, semantic or even discursive information.

5.6 Discussion

In this chapter we have provided empirical data that support our theoretical claims, from human judgements and from automatic procedures.

A corpus of texts summarized by different human judges has been studied, providing support to the claim that the proposed representation of discourse organization is useful to model the behaviour of human judges in fine-grained (intra-clausal) text summarization. Judges show a higher degree of agreement in discourse segments that occur in highly marked discourse relations, like *cause*, *revision* or *equality*. The probability that a word is removed, however, seems to be better modelled by shallow cues, like syntactic form or the presence of discourse markers. In a preliminary study we have shown that a combination of these two aspects, semantics and form, provides a good modelling of discourse structure for text summarization, identifying parts of text that judges tend to consider relevant or irrelevant, and parts where they tend to disagree.

A bigger corpus could give stronger statistical validity to the claim, and it could also be exploited to refine the conclusions about how different kinds of information (shallow cues, basic discursive semantics) influence the perception of relevance by human judges. To develop such corpus, texts have to be annotated with information about their surface properties and about the meaning of the relations between discourse units. The first kind of annotation can be carried out semiautomatically with the automatic procedures that have resulted from this thesis, and then validated by human judges. The second kind of annotation, however, is much more costly, because it involves a serious interpretation of texts, it even requires human judges to be instructed about the framework beforehand. As an alternative, existing corpora annotated with comparable discourse representations could also be exploited, like the RST corpus (Carlson, Marcu and Okurowski 2003). To do that, RST labels for relations should be mapped to our basic meanings.

We have also found that naive judges show a degree of agreement well beyond chance

when identifying our proposed semantics for relations between discourse segments, indicating that this representation captures speakers' intuitions. However, the agreement between judges does not reach reliable degrees of stability or reproducibility. This is probably due to the fact that judges are not trained professionals.

As a result of this annotation, some conclusions about the nature of relations between minimal discourse units have been drawn. In the first place, we have found that sentential syntax is almost always monotonically percolated to discourse structure. We have seen that the kind of relations studied have inter- and intra-sentential scope, but cannot capture relations between paragraphs. Finally, we have seen that discourse markers are highly informative of discourse structure, because judges hardly ever underspecify their meaning, but they tend to disagree on the discourse semantics they convey.

On the other hand, we have shown that our theoretical concepts can be well addressed by shallow techniques. We have presented two applications, one to identify discourse segments and another to identify new discourse markers. Studying these two applications we have shown which features are more useful to improve the performance of automatic tools.

We have presented a preliminary implementation of an algorithm for automatic segmentation of text, based in the theoretical concept presented in Section 3.3. We have presented results exploiting heterogeneous information at varying degrees of reliability, showing that discourse markers and punctuation are the most informative clues to identify discourse segments. We have compared results of automatic segmentation with two human-based gold standards, by precision and recall. Future work includes carrying out an extensive comparison between human and automatic procedures, identifying cases where human judges disagree, where judges agree but they disagree with automatic procedures and cases where both humans and automatic procedures agree.

We have also presented X-Tractor, a tool to create and most of all enhance discourse marker lexica based on the features characterizing discourse markers in huge amounts of text. We have shown that some features that are not strictly defining of discourse markers are nevertheless very useful to characterize them in corpus, for example the punctuation context, their discursive content, their internal structure, their lexic weight, etc.

It would be interesting to characterize the discourse markers in our prototypical lexicon using X-Tractor, and compare the performance of the tool with this smaller lexicon. It would also be interesting to compare the behaviour of parallel discourse markers in the three languages (Catalan, Spanish and English) when they are characterized by the features of their occurrences in corpus. Another interesting issue would be assessing the intuitions of judges with respect to what is and what is not a discourse marker. We could obtain measures of agreement for different human judges in the task of tagging the discourse marker candidates provided by X-Tractor with their degree of *markerhood*.

Then, the future work that most naturally follows from what has been exposed up to here is the implementation and evaluation of our model of discourse organization in an independent (not application-oriented) discourse parser. This implementation seems rather trivial given the hierarchy of markedness of discursive meanings presented in this thesis, and the attachment algorithm discussed in Appendix B.

Chapter 6

Conclusions

In this thesis we have addressed the problem of text summarization from a linguistic perspective. After reviewing some work in the area, we came to the conclusion that those approaches to text summarization that are satisfactory are precisely those that rely on general properties of language, and that these properties are reflected in the surface realization of texts. Based on this brief overview, my starting hypotheses were:

1. A representation of texts at discourse level is useful for AS systems to improve the quality of resulting summaries.
2. Such representation can be built based on evidence found at the surface realization of texts.

   (a) Systematic relations can be established between this evidence and the behaviour of human judges to summarize texts.

   (b) This evidence can be the basis to obtain (part of) the targeted representation of discourse by shallow NLP techniques, more concretely, by those techniques available for Catalan and Spanish.

In order to test the validity of these two main hypotheses, we have developed a framework to obtain a representation of texts at discourse level exploiting linguistic evidence at surface level. We have shown that this representation contributes to improve the quality of summaries in two different summarization approaches, and also that it is useful to model the behaviour of human judges in summarizing texts.

The work presented in this thesis does not suppose a giant step in the understanding of essential questions in the areas of automatic text summarization or discourse analysis. The main aim of this work is to constitute solid ground whereupon deeper, more insightful theories can be built, by systematizing and clarifying some properties of text that have been used in various approaches to text summarization in a rather unprincipled way.

6.1 Contributions

Laura, tú serás la generadora de obviedades.

Horacio Rodríguez, playing agent system game
Jaén 2001

We have provided a systematization of basic mechanisms of discourse organization that are useful for text summarization and can be captured with state-of-the art NLP techniques for Catalan and Spanish.

The aim of this systematization is to **explicitly relate surface features of texts that are indicative of discourse structure (*shallow cues*) with a theory of how discourse is organized**, always oriented to obtain a representation of texts that can improve the quality of automatic summaries. We have argued that this systematization is useful to overcome those limitations of shallow cues that lie only in the lack of principledness in the way they are exploited. Moreover, it is also useful to ground a theory of discourse organization that overcomes some of the controversies found in the literature, because of its strong empirical basis.

In order to carry out this systematization, we have determined **which shallow cues provide discursive information** within the scope of shallow NLP techniques, more concretely, within the capacities of NLP techniques currently available for Catalan and Spanish. Punctuation, some syntactical structures and, most of all, discourse markers, are especially useful to delimit minimal discourse units and the relations between them.

We determined that the information provided by this kind of clues is only reliable at a short scope, more concretely, at **inter- and intra-sentential scope**. Moreover, we also found that this information is often underspecified. Then, we specified a formal representation of the organization of discourse that captures the idiosincracies of the information obtained from shallow cues.

We propose that **discourse can be represented as a set of independent sequences of hierarchical trees**. Each of these sequences captures a different kind of relations between the same set of discursive units, so that heterogeneous meanings are explicitly separated, which provides an elegant way of capturing some configurations of relations that cannot be captured by strict tree-like representations. The fact that the structures are not a tree, but a sequence of trees, is determined by the limited scope of the information provided by the shallow cues we are exploiting. When a shallow cue is found, content-rich relations can be established between discourse units. But there are some parts of text where no such cues can be found and that are beyond the scope of surrounding shallow cues; in these cases, default relations are established between units. These default relations are expressed as precedence in a list.

Minimal discourse units have been defined explicitly relating their theoretical and empirical properties. We have distinguished two main kinds of units: those

conveying propositional content (segments) and those that signal how discourse should be organized (markers). We have developed automatic procedures to identify and characterize both kinds of units. On the one hand, we have evaluated the performance of different algorithms to automatically identify discourse segments in text, in terms of precision and recall as compared to two different gold standards produced by human judges. We have shown that punctuation and, most of all, discourse markers, are the most useful evidence to identify discourse segments in text. Then, we have applied lexical acquisition techniques to enhance a starting lexicon of discourse markers by exploiting surface features that characterize them in text. We have shown that also punctuation and also co-occurrence with other discourse markers are useful to identify previously unseen discourse markers in text.

An inventory of basic discursive meanings has been induced from the evidence provided by highly grammaticalized discourse markers, and a method for inducing it has been established. The meaning of relations between discourse units is described as a conglomerate of basic meanings. These basic meanings are organized in dimensions grouping heterogeneous meanings, which are in turn organized in a range of markedness that allows to assign an unmarked meaning to unmarked cases by default, thus guaranteeing that all relations between discourse units can be described within the presented framework, even without the presence of a shallow cue. We have induced a set of basic discursive meanings based on the evidence provided by discourse markers. In contrast to previous work pursuing this method (Knott 1996), we only exploit highly grammaticalized discourse markers, to identify basic distinctions in discursive meaning. Thus, we only provide a gross-grained description of the semantics of discourse relations, but with two interesting properties for NLP applications: that the set of basic meanings are much less controversial than in earlier approaches, and that they can be captured by shallow NLP techniques with a reliable degree of certainty.

We have carried out an experiment where three judges, two naive judges and a linguist, applied the proposed set of meanings to describe relations between discourse units. The level of agreement between judges is high enough to assert that **this set of basic meanings seems to capture speakers' intuitions about the organization of discourse**. Moreover, a closer study of the patterns of annotation shows that (a) discourse markers are highly informative of discourse structure, but their semantics is unclear, (b) the kind of relations studied have inter- and intra-sentential scope, but cannot capture relations between paragraphs and (c) that sentential syntax is percolated to discourse structure.

Finally, we have carried out some experiments showing that **the representation of discourse proposed in this thesis improves the quality of automatic summaries and is a good model of some aspects of human summarization**. We have integrated the proposed representation of discourse within two approaches to automatic summarization: combined with a lexical chain summarizer (Section 2.4.2.1) and as one of the analysis modules of an e-mail summarizer (Section 2.4.2.2). In both cases we have obtained that the discursive analysis contributed to the improvement of automatic summaries.

On the other hand, we have applied this analysis of discourse to model the behaviour of human judges in a task where they assessed the relevance of words in a text by removing those that they considered unnecessary. We have shown that the behaviour of judges

does not significantly differ from random if it is considered as it is, but when texts are represented with the proposed analysis of discourse, their behaviour can be described with much more accuracy: it is possible to identify parts of text where judges tend to agree, most of all, it is possible to identify parts of text that judges tend to consider irrelevant.

As a result of the work presented here, a number of **resources** have been created: annotated corpora, a lexicon of prototypical discourse markers and systematic procedures for identification of minimal discourse units and their relations, formalized as algorithms and, in some case, also implemented as an automatic procedure.

annotated corpora we have developed corpora for fine-grained evaluation of summarization, a small corpus of journalistic articles (described in Section 5.2, available at `http://lalonso.sdf-eu.org/compression/`) and a bigger corpus of e-mails (described in Section 2.4.2.2, available at `http://www.lsi.upc.es/~bcasas/carpanta/`). In another small corpus relations between discourse units have been identified and their meaning has been described (described in Section 5.3, available at `http://lalonso.sdf-eu.org` These three corpora are small, but provide empirical evidence for our theoretical claims, and have served as a testbed to establish annotation guidelines that can be used to enhance them in the future.

a lexicon of prototypical discourse markers containing 84 discourse markers with a near-parallel in Catalan, Spanish and English, characterized by the discursive meaning that they prototypically convey and by their syntactical behaviour (described in Appendix A). Being small, this lexicon has a very limited recall, but it can be enhanced applying X-Tractor, a lexical acquisition tool targeted to obtain previously unseen discourse markers from large amounts of corpus, with a good level of performance (described in Section 5.5).

systematic procedures for analysis of discourse organization we have developed an algorithm to identify discourse segments exploiting punctuation, some syntactic structures and discourse markers. Two different implementations have been evaluated, showing good performance. Then, an algorithm to determine attachment point and its topography, which, together with the organization of meaning in a range of markedness, make it trivial to implement an automatic procedure to identify and characterize relations between discourse units that have been previously identified by the segmentation algorithm.

6.2 Future Work

just for today...

Maria Fuentes, as quoting Laura Alonso, research stay Girona 2003

The work presented in this thesis is only a small step forward within the investigation of the systematic relations between shallow cues and theoretical approaches to the organization of texts. It builds mostly upon the work of Knott (1996), Marcu (1997b) and Forbes, Miltsakaki, Prasad, Sarkar, Joshi and Webber (2003), so most of the interesting lines of research are already suggested there. The final goal of all these approaches is to automatically obtain a comprehensive analysis of the discursive organization of texts, aiming to full natural language understanding. Of course we are still far from reaching that point, so the lines of future research we will discuss here are much less ambitious.

First of all, it would be interesting to provide a stronger empirical basis to the theoretical proposals of this thesis, enhancing the evaluation efforts presented in Chapter 5.

In Section 5.2, a corpus of texts summarized by different human judges was studied, providing support to the claim that the proposed representation of discourse organization is useful to model the behaviour of human judges in fine-grained, extractive text summarization. A bigger corpus could give stronger statistical validity to the claim, and it could also be exploited to refine the conclusions about how different kinds of information (shallow cues, basic discursive semantics) influence the perception of relevance by human judges.

However, in order to evaluate the kind of phenomena that we have studied in Section 5.2, texts have to be annotated with information about their surface properties and about the meaning of the relations between discourse units. The first kind of annotation can be carried out semiautomatically with the automatic procedures that have resulted from this thesis, and then validated by human judges. The second kind of annotation, however, is much more costly, because it involves a serious interpretation of texts, it even requires human judges to be instructed about the framework beforehand. As an alternative, existing corpora annotated with comparable discourse representations could also be exploited, like the RST corpus (Carlson, Marcu and Okurowski 2003). To do that, RST labels for relations should be mapped to our basic meanings.

But then, we require that texts have been summarized so that the kind of phenomena that can be modelled by our representation can be properly evaluated, that is, corpora where summarization has been carried out at a finer-grained level (intraclausal), thus excluding corpora summarized at sentence level. A corpus of pairs $< text, abstract >$ cannot be directly used, because we evaluate the relevance assigned to the effective words that can be found in a text, which is not straightforward if we cannot establish a direct mapping between words in the source text and in the summary. Some techniques have been proposed to establish this mapping (Knight and Marcu 2000; Jing 2001), but results

are approximate and would probably bias the result of the evaluation. So, besides the corpus presented in Section 5.2, the corpus that to our knowledge is most adequate to this kind of evaluation is that produced for evaluation of **Carpanta** (see Section 2.4.2.2). This corpus can be very useful to evaluate the adequacy of our model, as long as it is properly enriched with the kind of annotations presented in the previous paragraph.

In general, it would be interesting to compare the performance of our analysis of discourse and its contribution to the improvement of automatic summaries with similar approaches, for example, Marcu (1997b). It would also be interesting to investigate how the representation presented here interacts with summarization procedures other than the two that have been presented here.

Some of the automatic procedures presented in this thesis have been evaluated, as is the case of segmentation algorithms and the tool to identify previously unseen discourse markers. But some other procedures have not been implemented and therefore they have not been evaluated, as is the case of the algorithm to determine the attachment point of a given discourse segment and its topographical configuration, that is, the segment with which a discourse segment is related, and the structural properties of this relation. The procedure to label discourse relations with their semantic meaning has not been formalized in an algorithm, but its formalization and implementation is arguably trivial, given:

- the identification of discourse units,
- the identification of attachment points and their topographical configuration,
- the relation between surface cues (most of all discourse markers) and basic discursive meanings, and
- the organization of these meanings in a hierarchy of markedness, so that whenever two meanings are in conflict, the most marked is the one that remains.

Only cases of conflict between different amounts of evidence are still to be solved. The implementation and evaluation of stand-alone computer programs applying the procedures to determine attachment point and assign meaning to relations is left for future work.

Another interesting line of research would be to apply the presented methodology to induce basic discursive meanings to other languages, to test whether the proposed inventory of meanings can be considered basic cross-linguistically. Applying the methodology of semantic maps and taking into account only highly grammaticalized discourse markers, evidence from other languages can be easily integrated with what has been presented in this thesis, and corrected or refined if necessary. Once an inventory is well-established for a set of languages, basic lexica of discourse markers can be created for these languages, and then enhanced using X-Tractor.

These lexica would be the core of tools for automated identification of discourse units and their relations. It would be then possible to test whether the algorithms presented here for identification of segments and configuration of attachment point hold cross-linguistically or whether they are language-independent, as we suspect. To assess the performance of these automatic procedures, however, annotated corpora would have to be created. The effort required to build these corpora would be significantly reduced by pre-annotating

them with the automatic procedures and also because annotation guidelines are already well-established.

Then, a more ambitious, and quite obvious, line of future research consists in enhancing the scope of the presented theory of discourse organization, which is currently only able to reliably identify intra-sentential and inter-sentential discourse relations. The shallow cues that constitute the empirical basis of our theory of discourse do not seem to provide reliable information at long scope, then, it seems clear that, if the scope of the relations is to be enhanced in the representation of discourse, other kinds of information will have to be integrated. We made an initial exploration of how to combine the representation of discourse provided by our shallow cues with information provided by lexical chains, and obtained that the integration was successful, but had to be pursued further.

Appendix A

Lexicon of discourse markers

In this appendix we present the seminal discourse marker lexicon that we have used in this thesis. The discourse markers listed here were the primary source of evidence to draw the semantic maps to obtain an inventory of basic discursive meanings, presented in Section 4.2.3. This lexicon is also the basis for the implementations of a discourse segmenter (Section 5.4) and for the discourse analysis exploited by the e-mail summarizer Carpanta (Alonso, Casas, Castellón and Padró 2004b).

The lexicon is parallel in three languages: Catalan, Spanish and English. Therefore, in this starting version of the lexicon we have only included those discourse markers that have a near-synonym in one of the other languages. Those that do not have a near-synonym have been included in the extended version of the lexicon created by bootstrapping techniques applied to this starting lexicon (see Section 5.5).

As explained in Section 3.4, the discourse markers that constitute the prototypical lexicon were obtained from previous work, mostly Knott (1996) and Marcu (1997b), with the restriction that they are highly grammaticalized. We have also included in the lexicon some closed class words, obtained from the dictionary of the FreeLing morphosyntactic analyzer (FreeLing). We have discarded closed class words that are very vague (Table A.7) and highly ambiguous discourse markers (Table A.6).

In this lexicon, discourse markers are characterized by their structural (*continuation* or *elaboration*) and semantic (*revision, cause, equality, context*) meanings, and they are also associated to a morphosyntactic class (part of speech, PoS), one of *adverbial* (A), *phrasal* (P) or *conjunctive* (C) (see Section 3.4.3.2 for a description of these morphosyntactic classes). No information has been encoded about the reliability of discourse markers with respect to their discursive (vs. sentential) function. The only information of this kind that we provide is that discourse markers that are highly ambiguous with respect to their function (see Table A.7) are not included in the lexicon.

Sometimes a discourse marker is **underspecified** with respect to a meaning. We encode this with a hash. This tends to happen with structural meanings, because these meanings can well be established by discursive mechanisms other than discourse markers, and the presence of the discourse marker just reinforces the relation, whichever it may be.

Sometimes a discourse marker is **ambiguous** with respect to two meanings. In this cases, we write the predominant meaning in italics, and the secondary meaning in parentheses, or both of them in italics if no predominant meaning can be determined. Resolving

	revision	cause	equality	context	**total**
elaboration	4	9	10	22	41
continuation	9	9	6	4	28
underspecified	1	–	10	4	15
total	14	18	26	32	84

Table A.1: Distribution of the number of discourse markers accross the different meanings. Some discourse markers have been assigned to more or less than one meaning per dimension, because they are ambiguous or underspecified, respectively.

such ambiguities normally requires information about the context of occurrence, but we have not associated discourse markers with the contextual features that can be of aid to disambiguate them. Nevertheless, it seems that determining the adequate meaning associated to a particular instance of a discourse marker can be well addressed by general procedures, directly implemented in those algorithms that exploit the information stored in a lexicon (segmentation algorithms, discourse parsers, etc.).

All in all, the lexicon is formed by 84 discourse markers, representing different discursive meanings as can be seen in Table A.1.

We present the lexicon of prototypical discourse markers in Tables A.2 to A.6. Additionally, we also present some items that are not included in the seminal lexicon, but that have some relations with the items that are included: very vague closed class words are presented in Table A.7, and discourse markers without a parallel in some of the other languages are presented in Table A.8. Interesting aspects of the discourse markers in this lexicon are discussed after the corresponding tables. The lexicon can be found in electronic format, and enriched with illustrating examples for each of the discourse markers, at http://lalonso.sdf-eu.org/discmar/.

Catalan	**Spanish**	**English**	**structural**	**semantic**	PoS
a pesar de	despite	a pesar de	elaboration	revision	P
encara que	although	aunque	elaboration	revision	P
excepte	except	excepto	elaboration	revision	P
malgrat	in spite of	pese a	elaboration	revision	P
no obstant	however	no obstante	continuation	revision	A
nogensmenys	nevertheless	sin embargo	continuation	revision	A
en realitat	*actually*	en realidad	continuation	revision	A
de fet	*in fact*	de hecho	continuation	revision	A
al contrari	on the contrary	al contrario	continuation	revision	A
el fet és que	the fact is	el hecho es que	continuation	revision	P
és cert que	it is true that	es cierto que	continuation	revision	P
però	but	pero	continuation	revision	P
tot i això	even though	con todo	continuation	revision	A
ara bé	well now	ahora bien	continuation	revision	A
de tota manera	anyway	de todos modos	–	revision	A

Table A.2: Seminal discourse marker lexicon: discourse markers signalling revision.

however differs from *although* in their values for continuation or elaboration, although each of them can be used to rephrase the other in some contexts, *however* is attached to the segment that indicates continuation, *although* is attached to the segment that indicates elaboration.

actually / in fact their primary meaning is marking *evidentiality* (1), but they tend to be structurally equivalent to *however*, as we have shown using multiple alignment techniques (Alonso, Castellón, Escribano, Messeguer and Padró 2004c). In English their evidentiality meaning is more predominant than the revision meaning, and so their contribution as discourse markers of revision is only reliable when it co-occurs with other discourse markers also signalling revision (2-a) or in certain punctuation contexts (2-b), although they can also signal revision without any of these further evidence (3-a). In Spanish and Catalan their primary meaning is revision ((3-b) and (3-c)), respectively) comparable to *it is true that*. The kind of revision that these discourse markers tend to convey in Spanish and Catalan is *correction*[1]. We can speculate that the reason why the revision meaning of these discourse markers is more primary in Spanish or Catalan than in English is because in these languages the correction meaning tends to be expressed by discourse markers, as can be seen in the fact that it is lexicalized (*sinó, sino*), while in English it is covered by the all-purpose revision discourse marker *but*, and correction is only distinguished from other kinds of revision by other linguistic features.

[1]A prototypical example of correction would be: "This is not black, but white."

(1) They could also help themselves by thinking through a problem before phoning the support desk. In many cases a user will actually solve his or her own problem while on the phone to Neptune!

(2) a. Standardisation has never been the IT industry's strong point, and the answer is "probably not". However, they don't actually all do the same job.

b. He then argues that "it is not sufficient (for me) to tell the conference that there will be no return to mass picketing". Actually, I never mentioned picketing, mass picketing or otherwise, in my speech, but let that pass.

(3) a. Amnesty warmly welcomed the release of prisoners of conscience and the repeal of certain articles, but has urged that the legislation be extended to include reform or repeal of further articles of the Turkish Penal Code, under which POCs may be held.
The new law may in fact increase the already serious risk of torture facing political detainees.

b. La idea inicial de Maragall fue celebrar una exposición internacional, pero ese propósito falló cuando alguien de su gabinete descubrió que habían llegado tarde para obtener el reconocimiento internacional para un acontecimiento de este tipo. En realidad poco importaba qué se hiciera. Tanto Clos como Maragall perseguían en esencia poner una nueva fecha al futuro de la ciudad.

c. Tot va començar, com en les novel.les policíaques, amb un fiscal, entestat a treure a la llum el taló d'Aquil.les del president demòcrata. L'ham: una becària de 22 anys, grassoneta –usa la talla 46–, de pits exuberants i boca àmplia, una mica esbojarrada ja que creia tenir una relació sentimental quan en realitat va mantenir 10 trobades sexuals servides a domicili amb el senyor Clinton, qui, durant set mesos, es va obstinar a negar haver mantingut contacte físic amb ella.

it is true that in contrast with *actually* or *in fact*, its primary meaning is revision, like *en realidad, en realitat, de fet, de hecho* in Spanish and Catalan.

Catalan	**Spanish**	**English**	**structural**	**semantic**	PoS
donat que	given that	dado que	elaboration	cause	P
perquè	because	porque	elaboration	cause	P
degut a	due to	debido a	elaboration	cause	P
gràcies a	thanks to	gracias a	elaboration	cause	P
per si	in case	por si	elaboration	cause	P
per	because of	por	elaboration	cause	P
per això	that's why	por eso	continuation	cause	A
en conclusió	in conclusion	en conclusión	continuation	cause	A
així que	thus	así que	continuation	cause	P/A/P
com a conseqüència	as a consequence	como consecuencia	continuation	cause	A
per	in order to	para	continuation	cause	P
perquè	so that	para que	continuation	cause	P
per aquesta raó	for this reason	por esta razón	continuation	cause	A
per tant	so	por tanto	continuation	cause	A/C/A
en efecte	in effect	en efecto	continuation	cause	A

Table A.3: Seminal discourse marker lexicon: discourse markers signalling cause.

in conclusion while it looks similar to *in sum*, this discourse marker tends to convey new information, not to rephrase it. Compare the following example with the example for *in sum*. With respect to the effects on coherence and relevance, it is comparable to consecutive discourse markers like *that's why* or *so then*, which can also signal relations that are not motivated by a causal relation in the real world, but have the same rhetorical strength as those that are motivated by a real causal relation. It is comparable to *in effect*.

(4) The European Court further ruled in this case that Arts 48 and 59 of the EC Treaty do not prevent a member state from requiring that the exercise of the profession of auditor in that state by a person qualified to carry on that profession in another member state be conditions which are objectively necessary to guarantee observation of professional rules concerning the permanence of the infrastructure in place for the completion of the work, the effective presence in the member state and assurance of the observation of professional ethics, unless respect for such rules and conditions is already guaranteed by a reviseur d'entreprises, whether a natural person or a firm, established and recognised in the state, and in whose service is placed, for the duration of the work, the person who intends to exercise the profession of auditor. In conclusion, one has to wonder whether the borders are in fact open.

perquè / per in Catalan these discourse markers are underspecified with respect to structural meaning, they can be equivalent to *so that / to* (5-a) or to *because / because of* (5-b).

(5) a. Avui sento por perquè han declarat impunes tots els caps d'Estat.
Today I feel frightened because all heads of State have been declared impune.

b. La Generalitat ha fet una crida a la solidaritat perquè s'ocupin aquestes cases.
The Generalitat has made a call to solidarity so that these houses are occupied.

Catalan	**Spanish**	**English**	**structural**	**semantic**	PoS
en resum	in sum	en resumen	elaboration	equality	A
concretament	specifically	concretamente	elaboration	equality	A
en essència	essentially	en esencia	elaboration	equality	A
en comparació	in comparison	en comparación	elaboration	equality	A
en altres paraules	in other words	en otras palabras	elaboration	equality	A
en particular	in particular	en particular	elaboration	equality	A
és a dir	that is to say	es decir	elaboration	equality	C
per exemple	for example	por ejemplo	elaboration	equality	A
precisament	precisely	precisamente	elaboration	equality	A
tal com	such as	tal como	elaboration	equality	P
en darrer lloc	lastly	por último	continuation	equality	A
per una banda	on the one hand	por un lado	continuation	equality	A
per altra banda	on the other hand	por otro lado	continuation	equality	A
a propòsit	by the way	a propósito	continuation	equality	A
no només	not only	no sólo	continuation	equality	P
sinó també	but also	sino también	continuation	equality	P
en dues paraules	in short	en dos palabras	–	equality	A
a més	moreover	además	–	equality	A
també	also	también	–	equality	A
a banda	besides	aparte	–	equality	A
encara més	what's more	aún es más	–	equality	A
fins i tot	incluso	even	–	equality	P
especialment	specially	especialmente	–	equality	A
sobretot	above all	sobretodo	–	equality	A

Table A.4: Seminal discourse marker lexicon: discourse markers signalling equality.

not only ... but also

lastly unlike *first of all* or *to begin with*, and like *secondly, thirdly*, this discourse marker is not ambiguous, because it requires a context of sequence to be felicitous.

on the one hand / on the other hand like *lastly*, they require a sequence context to be felicitous, so they are not ambiguous with respect to their structural or semantic meaning, but their ambiguity with respect to scope varies greatly. If they co-occur (6), their scope can be determined if we consider that the scope of *on the one hand* reaches until the point of occurrence of *on the other hand*, and that the latter has a scope of an equivalent size. However, if *on the other hand* occurs alone, its scope is very hard to determine automatically, and probably also by human judges.

(6) It does occur to Fukuyama that religion might have some sort of unease to express with all this, but he appears to conceive of religion under only two modes. On the one hand, there is fundamentalist counter-ideology, the

Islamic theocratic state. This, it is to be assumed, his liberal readers may take seriously as a threat, but hardly as an option. And [on the other hand], there are "less organised religious impulses", religion as individual preference. This he knows can readily be accommodated another sort of consumer commodity, "within the sphere of personal life permitted in liberal societies".

in short is ambiguous with respect to continuation or elaboration, because the discourse unit to which the discoruse marker is attached can sometimes contribute new information, as in the following example.

(7) The authors maintain that the role of women in the Tigrayan society is still closely linked to their status in the feudal system. 1975 women were treated as children. They were not allowed to own land nor speak. [In short] women were at the bottom of the hierarchy of oppression with no rights of any kind.

in sum / essentially convey an elaborative relation because they repeats information that has already been given, even if this information is given in a shorter form. The utilitty of these discourse markers for automatic summarization is an *ad-hoc* property, subject to the task and not to their effects with respect to coherence and relevance assessment. Therefore, it has to be treated by manually creating a special rule that overrides general discursive rules.

Catalan	**Spanish**	**English**	**structural**	**semantic**	PoS
considerant	considering	teniendo en cuenta	elaboration	context	P
després	after	después	elaboration	context	P
abans	before	antes	elaboration	context	A
originalment	originally	originalmente	elaboration	context	A
a condició de	provided that	a condición de	elaboration	context	P
durant	during	durante	elaboration	context	P
mentre	while	mientras	elaboration	context	P
a no ser que	unless	a no ser que	elaboration	context	P
quan	when	cuando	elaboration	context	P
on	where	donde	elaboration	context	P
d'acord amb	in accordance with	de acuerdo con	elaboration	context	P
lluny de	far from	lejos de	elaboration	context	P
tan aviat com	as soon as	tan pronto como	elaboration	context	P
de moment	for the moment	por el momento	elaboration	context	A
entre	between	entre	elaboration	context	P
cap a	towards	hacia	elaboration	context	P
fins a	until	hasta	elaboration	context	P
mitjançant	by means of	mediante	elaboration	context	P
segons	following	según	elaboration	context	P
en qualsevol cas	in any case	en cualquier caso	continuation	context	A
aleshores	then	entonces	continuation	context	A
respecte de	with respect to	respecto a	continuation	context	P
en aquest cas	in that case	en ese caso	continuation	context	A
si	if	si	–	context	P
sempre que	whenever	siempre que	–	context	P
sens dubte	no doubt	sin duda	–	context	A
alhora	at the same time	a la vez	–	context	A

Table A.5: Seminal discourse marker lexicon: discourse markers signalling context.

first of all / to begin with as many discourse markers, these are lexically underspecified with respect to elaboration or continuation, they can reinforce progressive and elaborative relations that are actually signalled by means other than this discourse marker. They are also ambiguous between context and equality. If it is part of a sequence, as in example (8), it will signal equality, if not, as in example (8), it will signal context. By default, we ascribe it to context, and only if there is enough evidence is it ascribed to equality.

(8) a. Police say that cars are being stolen to be resold in car-starved east European countries. To begin with, thieves went for the likes of Golf GTis and BMWs, but now bread-and-butter cars are also being taken.

Catalan	**Spanish**	**English**	**structural**	**semantic**
com	*como*	like	elaboration	*equality, context, cause*
com	*como*	*since*	elaboration	*cause, context*
desde	desde	*since*	elaboration	*cause, context*
sobre	*sobre*	about	*continuation, elaboration*	context
sobre	*sobre*	over	*continuation, elaboration*	context
abans de res	first of all	antes que nada	–	*context* (or *equality*)
per començar	to begin with	para empezar	–	*context* (or *equality*)

Table A.6: Highly polysemic discourse markers.

b. In Four Saints Thomson's informality was given free reign since he first of all improvised the music at the piano then, when it stuck, wrote it down to a figured bass.

(9) a. But there never was a threat to a new German-American special relationship, since there never was such a special relationship to begin with.

b. "We are a bit of a way from that. But I certainly believe, first of all, we have to give what help we can," said Mr Hurd.

in any case indicates continuation and context. It seems to have effects comparable to revision, but it is hard to find what is denied. It seems that it has contrastive functions, which can be best attributed to the properties of continuation than to any possible revision. It is comparable to topic-based *but*, but in that case there seems to be more correlation with items signalling negative polarity, which seems to support an interpretation as revision. In this respect, it is different from *anyway*, which always conveys revision.

(10) In truth, however, humble photocopying has been overtaken by the wonders of the fax and personal computers complete with printers. Whatever the price of these latter (20 times their cost in the West) and reinforced customs procedures for their import, they are finding their way in. The controls in any case are surely doomed to fail.

then is characterized by the two least marked meanings in each dimension, which makes it very close to *narration*.

unless even if it has inherent negative polarity, it does not convey revision, but context, comparable to *if* or *in case*.

Catalan	Spanish	English
i	y/e	and
ni	ni	nor/neither
o	o/u	or
que	que	that
amb	con	with
sense	sin	without
contra	contra	against
en	en	in
a	at/to	a
	to	

Table A.7: Closed class words with very vague meaning.

Catalan	Spanish	English
sinó	sino	*correction*

Table A.8: Closed class words without a near synonym in some of the other two languages.

Appendix B

Heuristics to determine the structural configuration of discourse

As explained in Section 4.2.3.1.1, to obtain the topographical configuration of discourse structure, we have to determine, for each discourse unit:

a. its location in the structure of discourse, by determining the node in the structure (that is, the discourse unit) where it is attached
b. the topographical relation with the node where it is attached, that is, the shape of the arc linking the two nodes (discourse units)

In this Appendix we present the heuristics to address these two aspects of the configuration of discourse structure.

B.1 Heuristics to determine the most adequate attachment point

The problem of determining the attachment point of a discourse unit in a graph-like structure of discourse is a typical problem of formal discourse analysis. Polanyi (1988) and Webber (1988) modelled this problem considering discourse as an incremental hierarchical tree. In a strict tree-like structure, crossing branches are forbidden, so, a discourse unit can only be legally attached to the set of nodes that are found on a path from the root to the last node that has been attached to the structure. This set of nodes is known as the *right frontier* of the discourse tree, and they constitute the set of possible attachment points for incoming discourse units.

As explained in Section 3.2, we do not model discourse as a single tree, but as a sequence of trees. In this case, the right frontier contains the nodes that constitute the path from the root of the last local tree to the last node that has been attached to this tree. The only difference with the traditional concept of right frontier is that the number of possible attachment points will probably be smaller than for a discourse tree covering the whole text. But even if the set is smaller, very often, there will be more than one

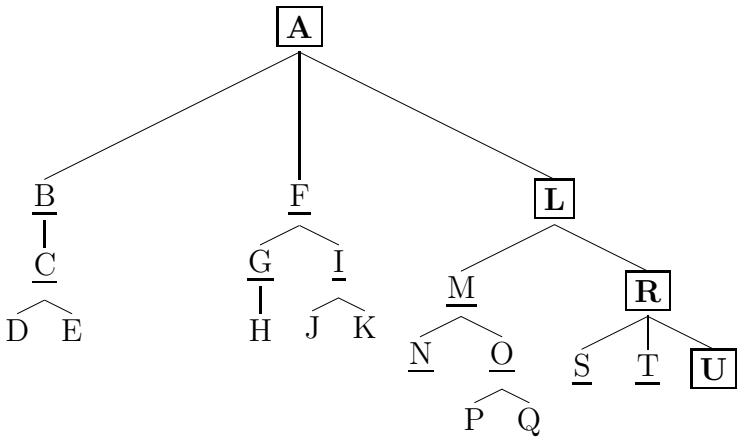


Figure B.1: Boxed nodes are those at the right frontier of discourse, underlined nodes are those at distance +2 from the right frontier, which we also take into consideration as possible attachment points. Alphabetic order indicates the order of occurrence of the segments in linear text.

possible attachment point for a discourse unit, and it will be necessary to determine the most adequate one.

Since the information provided by shallow techniques is error-prone, it is likely that a given representation of discourse does not adequately reflect the underlying structure. In an erroneous discourse structure, a discourse unit may not be attached to the most adequate node in the structure because it is not placed in the right frontier. Just like in *garden path* phenomena, an incoming discourse unit may provide information that indicates how to remake an inadequate representation so that it reflects the actual structure underlying the text.

Taking this into account, we enhance the set of possible attachment points to include nodes that are in a proximity of +2 nodes from the right frontier, as illustrated in Figure B.1. These nodes are also evaluated as possible attachment points, weighing their position in the structure as negative factor for the likelihood that the node is the most adequate attachment point for a given discourse unit. This allows that, if a node has been erroneously not placed at the right frontier and there is enough evidence that it is the attachment point for a discourse unit, the structure of discourse is remade to make this node part of the right frontier. This typically involves transforming a continuative relation (probably established by default) into an elaborative relation, as displayed in Figure B.2, where node E, having a continuative relation with node D, establishes an elaborative relation to allow F to be legally attached to D.

Many kinds of information can be of help to determine the best attachment point for a unit. Polanyi *et al.* (2004) claim that it can be satisfactorily determined exploiting sentential syntactic structures and lexical similarity. In our implementation, we determine the attachment point with a set of very simple heuristics.

In order to select the most adequate attachment, we first determine the subset of

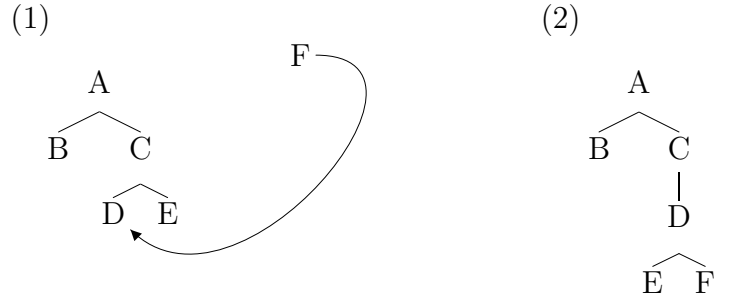


Figure B.2: Reconfiguration of a structure to accomodate the attachment of a discourse unit to a node that is not at the right frontier of discoruse, without crossing branches.

discourse units $A_1, A_2, ...A_n$ that are possible attachment points to a discourse unit DU. This subset is constituted by the nodes in the right frontier an those at distance +2 of the right frontier, as displayed in Figure B.1. Then, we rank the set of possible attachment points. For each A in $A_1, A_2, ...A_n$, we consider:

structural proximity of A to DU, so that nodes that are closer to the right frontier are ranked higher (examples of nodes referring to Figure B.1)

A is at the right frontier (`A, L, R, U`):	+10
A is at distance +1 of the right frontier(`B, F, M, S, T`):	+1
A is at distance +2 of the right frontier (`C, G, I, N, O`):	0

syntactical subordination of DU to A, so that if DU is subordinated to A, A is ranked higher

DU is a subordinate clause (relative, completive or non-personal) or phrase of A:	+10
DU is a subordinate clause of A dominated by a discourse marker:	+1

sequential proximity of A to DU, so that discourse units that have come recently before DU in the text are ranked higher

A is in the same sentence as DU, with 0 discourse units in between:	+10
A is in the same sentence as DU, with 1 discourse unit in between:	+5
A is in the same sentence as DU, with > 1 discourse units in between:	+1
A is in the sentence preceding DU, with 0 discourse units in between:	+5
A is in the sentence preceding DU, with 1 discourse unit in between:	+1

referential cohesion between A and DU, so that if DU contains some anaphoric expression whose referent is likely to be found in A, A is ranked higher

DU contains a relative pronoun and A is its matrix clause:	+10
DU contains a personal pronoun and A is the preceding clause:	+1

lexical similarity of A with respect to DU, so that discourse units that share more words with DU are ranked higher. The values for this feature are continuous, normalized to a scale ranging from 5 (sharing most words) to 0 (not sharing any word).

As a result of ranking all possible attachment points, we obtain an ordered list of possible attachment points, where the topmost A is the node where DU is most likely attached.

As explained above, if this topmost A is not in the right frontier, the topographical configuration of the structure of discourse must be changed, in order to allow that DU is attached to this A without crossing branches.

B.2 Heuristics to determine the topography of the attachment

Once the attachment point A for a discourse unit DU has been determined, the topography of the attachment of DU to A must be determined, configuring the shape of the resulting structure. This has implications with respect to the availability of nodes as attachment points for other discourse units.

As seen in Figure B.3, elaborative relations increase the number of nodes that are available for attachment of subsequent discourse units, that is, they enhance the right frontier. Continuative relations never add new nodes (B.4-d), and they may even reduce the number of nodes that are available for attachment, if the attachment point is in a high level of the structure (B.4-e). These effects apply to a restricted graph representation as the one we propose, as well.

These heuristics are the implementation of the semantics of the structural meaning of discourse relations, whose linguistic aspects are discussed in Section 4.2.3.1.1. The meaning of relations in the structural dimension determines the the shape of the edge linking the nodes (discourse units). Elaborative relations link a discourse unit DU at a lower level than its attachment point A, while continuative relations link them at the same level.

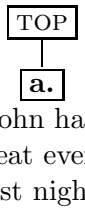
In case there is no shallow cue to determine the structural semantics of a relation, the default structural meaning, continuation, is applied. As a result, discourse units are attached in lineal order, without any hierarchical structure, as seen in Figure B.5. If a syntactical analysis is available, the default is to map sentential structure to discourse structure, as displayed in Figure B.6. When evidence from discourse markers is available, these default heuristics are enriched as shown in Figure B.6.

These heuristics can be enhanced to include cohesive links (lexical relatedness, coreference), document and genre structures, etc.

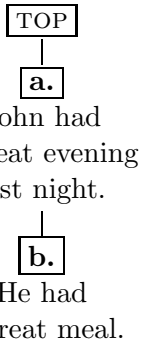
(1) **(a)**John had a great evening last night.
(b) He had a great meal.
(c) He ate salmon.
(d) He devoured lots of cheese.
(e) He then won a dancing competition.

adapted from Asher and Vieu (2001)

(a)



(b)



(c)

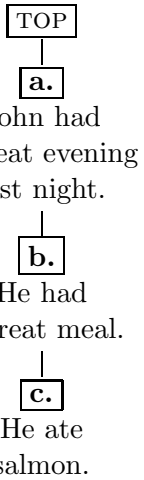
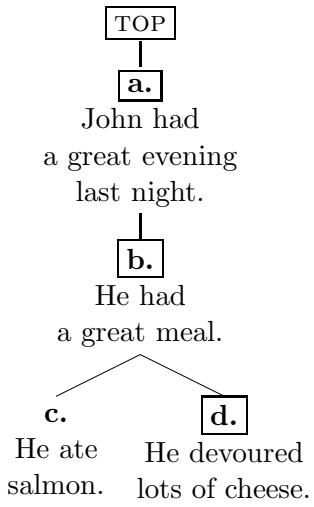


Figure B.3: Contribution of elaborative relations to shape the right frontier of a discourse structure, enhancing the number of nodes available for attachment (boxed nodes).

(2) **(a)**John had a great evening last night.
(b) He had a great meal.
(c) He ate salmon.
(d) He devoured lots of cheese.
(e) He then won a dancing competition.

adapted from Asher and Vieu (2001)

(d)



(e)

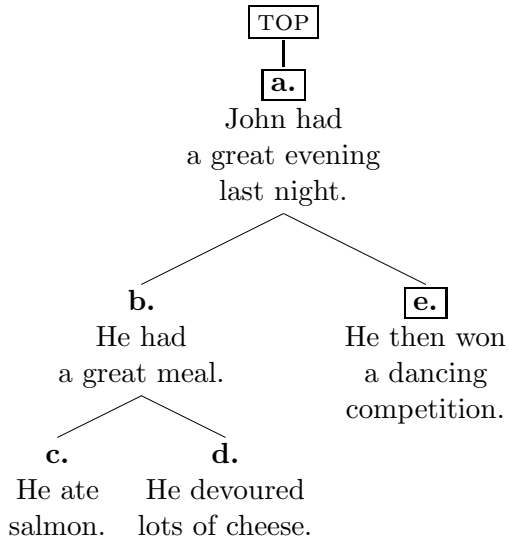


Figure B.4: Contribution of continuative relations to shape the right frontier of a discourse structure, not enhancing the number of nodes available for attachment (boxed nodes) or even reducing them.

(3) **a.**Roses are red,

b.violets are blue,

c.you love me

d.and I love you.

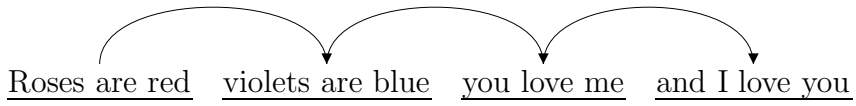


Figure B.5: Default representation of discourse as a sequence of discourse units related by the default structural relation, *continuation*, which is established by mere linear precedence of the units. This representation is obtained when no information about the configuration of discourse is available.

if *DU* is a main clause,
    DU is attached to *A*,
    A is not at the right frontier.
else
    DU is attached to *A*,
    both *A* and *DU* are at the right frontier.

Figure B.6: Heuristics to determine the structural meaning (or shape) of the discursive relation between a discourse unit *DU* and the node *A* where it is attached in the preceding structure of discourse. These heuristics are based on the syntactical structure of the sentence.

if *DU* is a main clause,

  if *DU* is dominated by a discourse marker indicating *continuation*,

    DU is attached to *A*,

    A is not at the right frontier.

  else

    if *DU* is dominated by a discourse marker indicating *elaboration*,

      DU is attached to *A*,

      both *A* and *DU* are at the right frontier.

    else

      DU is attached to *A*,

      A is not at the right frontier.

else

  if *DU* is dominated by a discourse marker indicating *continuation*,

    DU is attached to *A*,

    both *A* and *DU* are at the right frontier.

  else

    if *DU* is dominated by a discourse marker indicating *elaboration*,

      if *DU* precedes its main clause,

        DU is attached to *A*,

        DU is not at the right frontier.

      else

        DU is attached to *A*,

        both *DU* and *A* are at the right frontier.

    else

      DU is attached to *A*,

      A is not at the right frontier.

Figure B.7: Heuristics to determine the structural meaning (or shape) of the discursive relation between a discourse unit *DU* and the node *A* where it is attached in the preceding structure of discourse. These heuristics incorporate information provided by discourse markers.

Bibliography

Alonso, Laura (2001). Aproximació al Resum Automàtic per Marcadors Discursius. Master's thesis, Departament de Lingüística General, Universitat de Barcelona.

Alonso, Laura, Daniel Alonso, Ezequiel Andújar and Robert Sola (2004a). *Anotación discursiva de corpus en español. Tech. Rep.* Grial *report series 2004-2*, Departament de Lingüística General, Universitat de Barcelona.

Alonso, Laura, Bernardino Casas, Irene Castellón, Salvador Climent and Lluís Padró (2003a). Carpanta *eats words you don't need from e-mail.* In SEPLN, XIX Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural.

Alonso, Laura, Bernardino Casas, Irene Castellón, Salvador Climent and Lluís Padró (2003b). *Combining heterogeneous knowledge sources in e-mail summarization.* In Recent Advances in Natural Language Processing (RANLP 2003). Borovets, Bulgaria.

Alonso, Laura, Bernardino Casas, Irene Castellón and Lluís Padró (2004b). *Knowledge-intensive automatic e-mail summarization in* carpanta. In Proceedings of the 42nd meeting of the Association for Computational Linguistics. ACL'04. Demo session.

Alonso, Laura and Irene Castellón (2001). *Towards a delimitation of discursive segment for natural language processing applications.* In First International Workshop on Semantics, Pragmatics and Rhetoric. Donostia - San Sebastiàn.

Alonso, Laura, Irene Castellón, Salvador Climent, Maria Fuentes, Lluís Padró and Horacio Rodríguez (2003c). *Approaches to text summarization: Questions and answers.* In Revista Iberoamericana de Inteligencia Artificial, (20):pp. 34–52.

Alonso, Laura, Irene Castellón, Jordi Escribano, Xavier Messeguer and Lluís Padró (2004c). *Discovering discourse patterns by multiple sequence alignment.* In 4th International Conference on Language Resources and Evaluation (LREC 2004).

Alonso, Laura, Irene Castellón, Karina Gibert and Lluís Padró (2002a). *An empirical approach to discourse markers by clustering.* In CCIA, Congrés Català d'Intel·ligència Artificial. Springer-Verlag, Castelló.

Alonso, Laura, Irene Castellón, Lluís Padró and Karina Gibert (2002b). *Discourse marker characterisation via clustering: extrapolation from supervised to unsupervised corpora.* In SEPLN. Valladolid.

Alonso, Laura and Maria Fuentes (2002). *Collaborating discourse for text summarisation.* In Proceedings of the Seventh ESSLLI Student Session.

Alonso, Laura and Maria Fuentes (2003). *Integrating cohesion and coherence for text summarization*. In Proceedings of the EACL'03 Student Session, pp. 1–8. Budapest.

Alonso, Laura, Maria Fuentes, Horacio Rodríguez and Marc Massot (2004d). *Re-using high-quality resources for continued evaluation of automated summarization systems*. In 4th International Conference on Language Resources and Evaluation (LREC 2004).

Alonso, Laura, Jennafer Shih, Irene Castellón and Lluís Padró (2003d). *An analytic account of discourse markers for shallow NLP*. In Manfred Stede and Henk Zeevat, eds., The Meaning and Implementation of Discourse Particles, whorkshop at ESSLLI'03.

Amigó, Enrique, Julio Gonzalo, Victor Peinado, Anselmo Pe nas and Felisa Verdejo (2004). *An empirical study of information synthesis task*. In Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics.

Arévalo, Montse, Xavi Carreras, Lluís Màrquez, M.Antònia Martí, Lluís Padró and M.José Simón (2002). *A proposal for wide-coverage spanish named entity recognition*. In Procesamiento del Lenguaje Natural, vol. 1(3).

Asher, N. (1993). Reference to abstract objects in discourse. Kluwer Academic Publishers.

Asher, Nicholas and Alex Lascarides (2003). Logics of Conversation. Cambridge University Press.

Asher, Nicholas and Laure Vieu (2001). *Subordinating and coordinating discourse relations*. In Intl. Wkshp. on Semantic, Pragmatics and Rhetorics. San Sebastián.

Atserias, Jordi, Josep Carmona, Sergi Cervell, Lluís Màrquez, M. Antònia Martí, Lluís Padró, Roberto Placer, Horacio Rodríguez, Mariona Taulé and Jordi Turmo (1998a). *An environment for morphosyntactic processing of unrestricted spanish text*. In First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain.

Atserias, Jordi, Irene Castellón and Montse Civit (1998b). *Syntactic parsing of unrestricted spanish text*. In First International Conference on Language Resources and Evaluation. LREC, Granada.

Baldwin, Breck, Robert L. Donaway, Eduard H. Hovy, Elizabeth D. Liddy, Inderjeet Mani, Daniel Marcu, Kathleen R. McKeown, Vibhu O. Mittal, Marc Moens, Dragomir R. Radev, Karen Sparck Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel and Michael White (2000). *An evaluation road map for summarization research*. TIDES.

Ballard, D., R. Conrad and R. Longacre (1971). *The deep and surface grammar of interclausal relations*. In Foundations of Language, (4):pp. 70–118.

Barzilay, Regina (1997). Lexical Chains for Summarization. Master's thesis, Ben-Gurion University of the Negev.

Barzilay, Regina, Noemie Elhadad and Kathy McKeown (2001). *Sentence ordering in multidocument summarization*. In HLT'01.

Barzilay, Regina, Kathy McKeown and Michel Elhadad (1999). *Information fusion in the context of multi-document summarization*. In Proceedings of ACL 1999.

Benitez, A. B. and S.-F. Chang (2002). *Multimedia knowledge integration, summarization and evaluation*. In Proceedings of the 2002 International Workshop On Multimedia Data Mining in conjuction with the International Conference on Knowledge Discovery and Data Mining (MDM/KDD-2002). Edmonton, Alberta.

Boguraev, Branimir K., Rachel Bellamy and Calvin Swart (2001). *Summarisation miniaturisation: Delivery of news to hand-helds*. In NAACL'01.

Boguraev, Branimir K. and Mary S. Neff (2000). *Lexical cohesion, discourse segmentation and document summarization*. In RIAO. Paris.

Bordería, Salvador Pons (1998). Conexión y Conectores: Estudio de su relación en el registro informal de la lengua. Cuadernos de Filología, Universitat de València.

Brunn, Meru, Yllias Chali and Christopher J. Pinchak (2001). *Text Summarization using lexical chains*. In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001. New Orleans, Louisiana.

Buyukkokten, Orkut, Hector Garcia-Molina and Andreas Paepcke (2001). *Text summarization of web pages on handheld devices*. In NAACL'01.

Carletta, Jean (1996). *Assessing agreement on classification tasks: the kappa statistic*. In Computational Linguistics, vol. 22(2):pp. 249–254.

Carletta, Jean, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwnyeth Doherty-Sneddon and Anne Anderson (1996). *Hcrc dialogue structure coding manual. Tech. Rep. HCRC/TR-82*, HCRC.

Carlson, Lynn and Daniel Marcu (2001). *Discourse tagging manual. Tech. Rep. ISI-TR-545*.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski (2001). *Building a discourse-tagged corpus in the framework of rhetorical structure theory*. In 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001. SIGDIAL.

Carlson, Lynn, Daniel Marcu and Mary Ellen Okurowski (2003). *Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory*. In Jan van Kuppevelt and Ronnie Smith, eds., Current Directions in Discourse and Dialogue. Kluwer Academic Publishers.

Chen, Hsin-Hsi (2002). *Multilingual summarization and question answering*. In Workshop on Multilingual Summarization and Question Answering (COLING'2002).

Chen, Hsin-Hsi, June-Jei Kuo and Tsei-Chun Su (2003). *Clustering and visualization in a multi-lingual multi-document summarization system*. In Proceedings of the 25th European Conference on IR Research, pp. 266–280.

Climent, Salvador, P. Gispert-Saüch, J. Moré, A. Oliver, M. Salvatierra, I. Sànchez, M. Taulé and Ll. Vallmanya (2003). *Machine translation of newsgroups at the uoc. evaluation and settings for language control*. In Journal of Computer-Mediated Communication. In press.

Cohen, J. (1960). *A coefficient of agreement for nominal scales*. In Educational & Psycological Measure, vol. 20:pp. 37–46.

Cooper, Robin, Staffan Larsson, Colin Matheson, Massimo Poesio and David Traum (1999). *Coding instructional dialogue for information states. Tech. Rep. D1.1*, The TRINDI Consortium.

Core, Mark G. and James F. Allen (1997). *Coding dialogs with the damsl annotation scheme*. In Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines.

Corston-Oliver, Simon H. (1998). Computing representations of the structure of written discourse. Ph.D. thesis, University of California, Santa Barbara.

Corston-Oliver, Simon H. (2001). *Text compaction for display on very small screens*. In NAACL'01.

Corston-Oliver, Simon H. and W. Dolan (1999). *Less is more: Eliminating index terms from subordinate clauses*. In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), pp. 348 – 356.

Cresswell, Cassandre, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber (2002). *The discourse anaphoric properties of connectives*. In 4th Discourse Anaphora and Anaphor Resolution Colloquium.

Croft, W. A. (2001). Radical Construction Grammar. Syntactic Theory in Typological Perspective. Oxford University Press.

Daumé III, H., Abdessamad Echihabi, Daniel Marcu, D.S. Munteanu and Radu Soricut (2002). *GLEANS: A generator of logical extracts and abstracts for nice summaries*. In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Di Eugenio, Barbara, Johanna D. Moore and Massimo Paolucci (1997). *Learning features that predict cue usage*. In ACL-EACL97, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 80–87. Madrid, Spain.

DiEugenio, B. and M. Glass (2004). *The kappa statistic: A second look*. In Computational Linguistics, vol. 30(1):pp. 95–101.

Discourse Resource Initiative (1997). *Standards for dialogue coding in natural language processing. Tech. Rep. 167*, Dagstuhl-Seminar.

DUC (2005). http://duc.nist.gov.

Edmunson, H. P. (1969). *New methods in automatic extracting*. In Journal of the Association for Computing Machinery, vol. 16(2):pp. 264 – 285.

Elhadad, Michael and Kathleen R. McKeown (1988). *What do you need to produce a 'but'? Tech. Rep. CUCS-334-88*, Department of Computer Science, Columbia University, New York.

Elhadad, Michael and Kathleen R. McKeown (1990). *Generating connectives*. In COLING 1990, vol. 3, pp. 97–102.

Elhadad, Noemie and Kathleen R. McKeown (2001). *Towards generating patient specific summaries of medical articles*. In NAACL'01 Automatic Summarization Workshop.

Endres-Niggemeyer, Brigitte (1998). Summarizing information. Springer.

Endres-Niggemeyer, Brigitte, Jerry Hobbs and Karen Sparck-Jones (1993). *Summarizing Text for Intelligent Communication*. Schloss Dagstuhl, Wadern, Germany. Dagstuhl Seminar Report IBFI GmbH.

Engehard, C. and L. Pantera (1994). *Automatic natural acquisition of a terminology*. In Journal of Quantitative Linguistics, vol. 2(1):pp. 27–32.

Fais, L. and K. Ogura (2001). *Discourse issues in the translation of japanese e-mail*. In Proceedings of the Pacific Association for Computational Linguistics, PACLING 2001.

Ferrara, K., H. Brunner and G. Whittemore (1990). *Interactive written discourse as an emergent register*. In Written Communication, vol. 8:pp. 8–34.

Forbes, K., C. Creswell, E. Miltsakaki, R. Prasad, Aravind K. Joshi and Bonnie Lynn Webber (2002). *The discourse anaphoric properties of connectives*. In DAARC.

Forbes, K., E. Miltsakaki, R. Prasad, A. Sarkar, Aravind K. Joshi and Bonnie Lynn Webber (2003). *D-LTAG system - discourse parsing with a lexicalized tree-adjoining grammar*. In Journal of Language, Logic and Information.

FreeLing (). `http://www.lsi.upc.es/~nlp/freeling/`.

Fuentes, Maria, Marc Massot, Horacio Rodríguez and Laura Alonso (2003). *Headline extraction combining statistic and symbolic techniques*. In DUC03. Association for Computational Linguistics, Edmonton, Alberta, Canada.

Fuentes, Maria and Horacio Rodríguez (2002). *Using cohesive properties of text for automatic summarization*. In JOTRI'02.

Givón, Talmy, ed. (1983). Topic continuity in discourse: a quantitative cross-language study. John Benjamins.

Gladwin, Philip, Stephen Pulman and Karen Sparck-Jones (1991). *Shallow Processing and Automatic Summarizing: A First Study. Tech. Rep. Technical Report No. 223*, University of Cambridge Computer Laboratory.

Goldstein, Jade, Vibhu O. Mittal, Mark Kantrowitz and Jaime G. Carbonell (1999). *Summarizing text documents: Sentence selection and evaluation metrics*. In SIGIR-99.

Grice, H. Paul (1969). *Utterer's meaning and intentions*. In Philosophical Review, vol. 68(2):pp. 147–177.

Grimes, Joseph E. (1975). *The Thread of Discourse*. In Jangua Linguarum, Series Minor, (207).

Grosz, Barbara J. and Candace L. Sidner (1986). *Attention, Intention, and the Structure of Discourse*. In Computational Linguistics, vol. 12(3):pp. 175–204.

Hahn, Udo (1990). *Topic Parsing: Accounting for Text Macro Structures in Full-Text Analysis*. In Information Processing & Management, vol. 26(1):pp. 135–170.

Hahn, Udo and Inderjeet Mani (2000). *The cahllenges of automatic summarization*. In IEEE Computer, vol. 33(11):pp. 29–36.

Halliday, M. A. K. and Ruqaiya Hasan (1976). Cohesion in English. English Language Series. Longman Group Ltd.

Harabagiu, S.M. and Finley Lacatusu (2002). *Generating single and multi-document summaries with GISTEXTER*. In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Haspelmath, Martin (2003). *The geometry of grammatical meaning: semantic maps and cross-linguistic comparison*. In Michael Tomasello, ed., The new psychology of language, vol. II, pp. 211–243. Lawrence Erlbaum, New York.

Hatzivassiloglou, Vassileios, Judith L. Klavans and Eleazar Eskin (1999). *Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning*. In EMNLP/VLC'99. Maryland.

Hatzivassiloglou, Vassileios, Judith L. Klavans, M. Holcombe, Regina Barzilay, M.Y. Kan and K.R. McKeown (2001). *Simfinder: A flexible clustering tool for summarization*. In NAACL'01 Automatic Summarization Workshop.

Hauptmann, A. G. and M. J. Witbrock (1997). *Informedia: News-on-demand multimedia information acquisition and retrieval*. In Mark T. Maybury, ed., Intelligent Multimedia Information Retrieval, pp. 215–239. AAAI/MIT Press.

Hearst, Marti A. (1994). *Multi-paragraph segmentation of expository text*. In 32nd Annual Meeting of Association for Computational Linguistics.

Herring, S. (1999). *Interactional coherence in cmc*. In Journal of Computer-Mediated Comunication, vol. 4(4). Special issue on Persistent Conversation.

Hirschberg, Julia and Barbara Grosz (). *Intonational features of local and global discourse structure*. In Speech and Natural Language Workshop. Morgan Kaufmann.

Hirschberg, Julia and Diane J. Litman (1993). *Empirical studies on the disambiguation of cue phrases*. In Computational Linguistics, vol. 19(3):pp. 501–529.

Hobbs, J. R. (1978). *Why is discourse coherent. Tech. Rep. technical note 176*, SRI International, Artificial Intelligence Center.

Hobbs, Jerry R. (1985). *On the coherence and structure of discourse. Tech. Rep. CSLI–85–37*, Center for the Study of Language and Information, Stanford University, Calif., USA.

Hovy, Eduard H. (1988). *Planning coherent multisentential text*. In ACL 1988, pp. 163–169.

Hovy, Eduard H. (2001). Handbook of Computational Linguistics, chap. 28: Text Summarization. Oxford University Press.

Hovy, Eduard H. and Chin-Yew Lin (1999). *Automated Text Summarization in SUMMARIST*. In Mani and Maybury, eds., Advances in Automatic Text Summarization.

Hovy, Eduard H. and Elisabeth A. Maier (1995). *Parsimonious or profligate: How many and which discourse structure relations?* Unpublished ms.

Hovy, Eduard H. and Daniel Marcu (1998). *Automated Text Summarization*. COLING-ACL. Tutorial.

Hutchinson, Ben (2004). *Acquiring the meaning of discourse markers*. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp. 685–692.

Jing, Hongyan (2001). Cut-and-Paste Text Summarization. Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University.

Jing, Hongyan and Kathleen R. McKeown (2000). *Cut and paste based text summarization*. In 1st Conference of the North American Chapter of the Association for Computational Linguistics.

Kageura, Kyo and Bin Umino (1996). *Methods of automatic term recognition: A review*. In Terminolgy, vol. 3(2):pp. 259–289.

Kan, Min-Yen (2003). Automatic text summarization as applied to information retrieval: Using indicative and informative summaries. Ph.D. thesis, Columbia University.

Kan, Min-Yen, Judith L. Klavans and Kathleen R. McKeown (2001). *Domain-specific informative and indicative summarization for information retrieval*. In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001. New Orleans.

Kan, Min-Yen and Kathleen R. McKeown (1999). *Information extraction and summarization: Domain independence through focus types. Tech. rep.*, Computer Science Department, Columbia University, New York.

Kehler, Andrew (2002). Coherence, Reference, and the Theory of Grammar. CSLI Publications.

Kintsch, Walter (1988). *The role of knowledge in discourse comprehension: A construction-integration model*. In Psychological Review, vol. 2(95):pp. 163–182.

Kintsch, Walter and Teun A. van Dijk (1983). Strategies of Discourse Comprehension. Academic Press.

Knight, Kevin and Daniel Marcu (2000). *Statistics-based summarization - step one: Sentence compression*. In The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000. Austin, Texas.

Knott, Alastair and Robert Dale (1996). *Choosing a Set of Coherence Relations for Text Generation: a Data-Driven Approach*. In .

Knott, Alistair (1996). A Data-Driven Methodology for Motivating a Set of Coherence Relations. Ph.D. thesis, University of Edinburgh, Edinburgh.

Knott, Alistair, Jon Oberlander, Mick O'Donnell and Chris Mellish (2001). *Beyond elaboration: The interaction of relations and focus in coherent text*. In Ted J. M. Sanders, Joost Schilperoord and Wilbert P. M. Spooren, eds., Text representation: linguistic and psycholinguistic aspects, pp. 181–196. Benjamins.

Kortmann, B. (1997). Adverbial subordination: a typology and history of adverbial subordination based on European languages. Berlin.

Kozima, Hideki (1993). *Text segmentation based on similarity between words.* In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics, pp. 286–288.

Kraaij, Wessel, Martin Spitters and Anette Hulth (2002). *Headline extraction based on a combination of uni- and multidocument summarization techniques.* In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Krippendorf, Klaus (1980). Content analysis: an introduction. Sage, Beverly Hills, California.

Kupiec, Julian, Jan O. Pedersen and Francine Chen (1995). *A trainable document summarizer.* In Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68–73. ACM Press.

Lagerwerf, Luuk (1998). Causal Connectives Have Presuppositions; Effects on Coherence and Discourse Structure. Den Haag, Holland Academic Graphics.

Lal, P. and S. Rueger (2002). *Extract-based summarization with simplification.* In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Landis, J. Richard and Gary G. Koch (1977). *The measurement of observer agreement for categorical data.* In Biometrics, vol. 33:pp. 159–174.

Langacker, Ronald W. (1987). Foundations of Cognitive Grammar, vol. I: Theoretical Prerequisites. University Press, Stanford.

Lascarides, Alex and Nicholas Asher (1993). *Temporal interpretation, discourse relations and commonsense entailment.* In Linguistics and Philosophy, vol. 16(5):pp. 437–493.

Lasswell, Steven (1996). *The alternation of verbal mood in Yup'ik Eskimo narrative.* In Marianne Mithun, ed., Prosody, grammar and discourse in Central Alaskan Yup'ik, pp. 64–97. University of California, Department of Linguistics, Santa Barbara.

Laura Alonso i Alemany, Ezequiel Andújar Hinojosa and Robert Sola Salvatierra (2004). *A framework for feature-based description of low level discourse in discourse annotation.* In Bonnie Webber and Donna Byron, eds., Workshop on Discourse Annotation. ACL'04.

Lehmam, Abderrafih and Philippe Bouvet (2001). *Évaluation, rectification et pertinence du résumé automatique de texte pour une utilisation en réseau.* In S. Chaudiron and C. Fluhr, eds., III Colloque d'ISKO-France: Filtrage et résumé automatique de l'information sur les réseaux.

Leuski, Anton, Chin-Yew Lin and Eduard H. Hovy (2003). *iNeATS: Interactive multi-document summarization.* In ACL'03.

Lin, Chin-Yew (1998). *Assembly of Topic Extraction Modules in SUMMARIST.* In Eduard H. Hovy and Dragomir R. Radev, eds., Proceedings of the AAAI Symposium on Intelligent Text Summarization, pp. 34–43. The AAAI Press, Stanford, California, USA.

Lin, Chin-Yew (2001). *Summary Evaluation Environment.* Http://www.isi.edu/~cyl/SEE.

Lin, Chin-Yew (2004). http://www.isi.edu/~ cyl/ROUGE/.

Lin, Chin-Yew and Eduard H.Hovy (1997). *Identifying topics by position.* In Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP97).

Lin, Chin-Yew and Eduard Hovy (2002a). *Manual and Automatic Evaluation of Summaries.* In Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Lin, Chin-Yew and Eduard Hovy (2002b). *NeATS in DUC 2002.* In Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Lin, Chin-Yew and Eduard Hovy (2003). *Automatic evaluation of summaries using n-gram co-occurrence statistics.* In Marti A. Hearst and Mari Ostendorf, eds., HLT-NAACL 2003: Main Proceedings, pp. 150–157. Association for Computational Linguistics, Edmonton, Alberta, Canada.

Lin, Chin-Yew and Eduard H. Hovy (2002c). *NeATS in DUC 2002.* In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Litman, Diane J. (1996). *Cue phrase classification using machine learning.* In Journal of Artificial Intelligence Research, vol. 5:pp. 53–94.

Longacre, Robert E. (1983). The Grammar of Discourse: Notional and Surface Structures. Plenum Press, New York.

Luhn, H. P. (1958a). *The automatic creation of literature abstracts.* In IBM Journal of research and development, vol. 2(2):pp. 159 – 165.

Luhn, H. P. (1958b). *The Automatic Creation of Literature Abstracts.* In IBM Journal of Research Development, vol. 2(2):pp. 159–165.

Malchukov, Andrej (2004). *Towards a semantic typology of adversative and contrast marking.* In Journal of Semantics, vol. 21(2):pp. 177–198.

Mani, Inderjeet (2001). Automatic Summarization. Nautral Language Processing. John Benjamins Publishing Company.

Mani, Inderjeet and Eric Bloedorn (1999). *Summarizing similarities and differences among related documents.* In Information Retrieval, vol. 1(1-2):pp. 35–67.

Mani, Inderjeet and Mark T. Maybury, eds. (1999). Advances in automatic text summarization. MIT Press.

Mann, William C. and Sandra A. Thompson (1988). *Rhetorical structure theory: Toward a functional theory of text organisation.* In Text, vol. 3(8):pp. 234–281.

Marcu, Daniel (1997a). *From discourse structures to text summaries.* In Mani and Maybury, eds., Advances in Automatic Text Summarization, pp. 82 – 88.

Marcu, Daniel (1997b). The Rhetorical Parsing, Summarization and Generation of Natural Language Texts. Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

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

Marcu, Daniel, Hal Daumé, Abdessamad Echihabi, Dragos Stefan Munteanu and Radu Soricut (2002). *GLEANS: A Generator of Logical Extracts and Abstracts for Nice Summaries*. In Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Martin, J. (1992). English Text: System and Structure. Benjamin, Amsterdam.

Martín Zorraquino, M. Antonia and José Portolés (1999). *Los marcadores del discurso*. In Ignacio Bosque and Violeta Demonte, eds., Gramática Descriptiva de la Lengua Española, vol. III, pp. 4051–4213. Espasa Calpe, Madrid.

Maybury, Mark T. and Inderjeet Mani (2001). *Automatic summarization*. ACL/EACL'01. Tutorial.

Maybury, Mark T. and Andrew E. Merlino (1997). *Multimedia summaries of broadcast news*. In International Conference on Intelligent Information Systems.

McKeown, Kathleen R., David Kirk Evans, A. Nenkova, Regina Barzilay, Vasileios Hatzivassiloglou, Barry Schiffman, Sasha Blair-Goldensohn, Judith L. Klavans and S. Sigelman (2002). *The columbia multi-document summarizer for DUC 2002*. In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

McKeown, Kathleen R., Judith L. Klavans, Vassileios Hatzivassiloglou, Regina Barzilay and Eleazar Eskin (1999). *Towards multidocument summarization by reformulation: Progress and prospects*. In AAAI 99.

McKeown, Kathleen R. and Dragomir R. Radev (1995). *Generating summaries of multiple news articles*. In ACM Conference on Research and Development in Information Retrieval SIGIR'95. Seattle, WA.

MEAD (). *Meadeval*. http://perun.si.umich.edu/clair/meadeval/.

Merchant, Jason (2003). *Fragments and ellipsis*. In Small structures: sentential and nonsentential analyses. Wayne University. Fall 2003 Colloquium Series.

Miller, G. A., R. Beckwith, Christiane Fellbaum, D. Gross, K. Miller and R. Tengi (1991). *Five papers on wordnet*. In Special Issue of the International Journal of Lexicography, vol. 3(4):pp. 235–312.

Mimouni, Naila K. (2003). *Segmentation for rhetorical representation purposes*. In RANLP'03.

MINIPAR (1998). `www.cs.ualberta.ca/~lindek/minipar.htm`.

Moore, Johanna D. and Martha E. Pollack (1992). *A problem for RST: the need for multi-level discourse analysis*. In Computational Linguistics, vol. 18(4):pp. 537–544.

Morris, Jane and Graeme Hirst (1991). *Lexical cohesion, the thesaurus, and the structure of text*. In Computational linguistics, vol. 17(1):pp. 21–48.

Moser, M. G., Johanna D. Moore and E. Glendening (1996). *Instructions for coding explanations: Identifying segments, relations and minimal units. Tech. Rep. Technical Report 96-17*, University of Pittsburgh, Department of Computer Science.

Muresan, S., E. Tzoukermann and Judith L. Klavans (2001). *Combining linguistic and machine learning techniques for email summarization*. In ACL-EACL'01 CoNLL Workshop.

Murray, D. E. (2000). *Protean communication: the language of computer-mediated communication*. In Tesol Quarterly, vol. 34(3):pp. 397–421.

Nenkova, Ani and Rebecca Passonneau (2004). *Evaluating content selection in summarization: the pyramid method*. In NAACL-HLT 2004.

Oates, Sarah Louise (2001). *A listing of discourse markers. Tech. Rep. ITRI-01-26*, Information Technology Research Institute, University of Brighton.

Ono, Kenji, Kazuo Sumita and Seiji Miike (1994a). *Abstract generation based on rhetorical structure extraction*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), pp. 344 – 348. Kyoto, Japan.

Ono, Kenji, Kazuo Sumita and Seiji Miike (1994b). *Abstract generation based on rhetorical structure extraction*. In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), pp. 344 – 348. Kyoto, Japan.

Paice, Chris D. (1990). *Constructing literature abstracts by computer*. In Information Processing & Management, vol. 26(1):pp. 171 – 186.

Palomar, M., A. Ferrández, L. Moreno, P. Martínez-Barco, J. Peral, M. Saiz-Noeda and R. Muñoz (2001). *An algorithm for anaphora resolution in spanish texts*. In Computational Linguistics, vol. 27(4).

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (2001). *BLEU: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176*, IBM.

Pardo, T.A.S., L.H.M. Rino and M.G.V. Nunes (2003). *GistSumm: A summarization tool based on a new extractive method*. In N.J. Mamede, J. Baptista, I. Trancoso and M.G.V. Nunes, eds., 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken, no. 2721 in Lecture Notes in Artificial Intelligence, pp. 210–218. Springer-Verlag.

Passonneau, Rebecca J. and Diane J. Litman (1997a). *Discourse segmentation by human and automated means*. In Computational Linguistics, vol. 23(1):pp. 103 – 140.

Passonneau, Rebecca J. and Diane J. Litman (1997b). *Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices*. In Eduard Hovy and Donia Scott, eds., Interdisciplinary Perspectives on Discourse. Springer Verlag.

Polanyi, Livia (1988). *A formal model of the structure of discourse*. In Journal of Pragmatics, vol. 12:pp. 601–638.

Polanyi, Livia (1996). *The lingustic structure of discourse. Tech. Rep. 96-200*, CSLI.

Polanyi, Livia, Chris Culy, Martin van den Berg, Gian Lorenzo Thione and David Ahn (2004). *A rule based approach to discourse parsing*. In SIGDIAL 2004.

Portolés, José (1998). Marcadores del discurso. Ariel.

Potsdam Corpus (2004). *The Potsdam commentary corpus*. `http://www.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.en.html`.

Quinlan, John R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Radev, Dragomir R. (2000). *Text Summarization*. ACM SIGIR. Tutorial.

Radev, Dragomir R., Weiguo Fan and Zhu Zhang (2001). *Webinessence: A personalized web-based multi-document summarization and recommendation system*. In NAACL Workshop on Automatic Summarization. Pittsburgh.

Radev, Dragomir R., Jahna Otterbacher, Hong Qi and Daniel Tam (2003). *MEAD ReDUCs: Michigan at DUC 2003*. In DUC03. Association for Computational Linguistics, Edmonton, Alberta, Canada.

Reitter, David (2003a). Rhetorical analysis with rich-feature support vector models. Master's thesis, University of Potsdam.

Reitter, David (2003b). *Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models*. In Uta Seewald-Heeg, ed., Sprachtechnologie für die multilinguale Kommunikation. St. Augustin, Germany: Gardez. URL `http://www.reitter-it-media.de/compling/papers/reitter_complex-rst_2003.pdf`.

Salton, Gerard, Amit Singhal, Mandar Mitra and Chris Buckley (1997). *Automatic text structuring and summarization*. In Information Processing and Management, vol. 33(3):pp. 193 – 207.

Sanders, Ted J. M. and Wilbert P. M. Spooren (2001). *Text representation as an interface between language and its users*. In T. Sanders, J. Schilperoord and W. Spooren, eds., Text representation: Linguistic and psycholinguistic aspects. Benjamins, Amsterdam.

Sanders, Ted J. M., Wilbert P. M. Spooren and Leo G. M. Noordman (1992). *Toward a taxonomy of coherence relations*. In Discourse Processes, vol. 15:pp. 1–35.

Santos Río, Luis (2003). Diccionario de Partículas. Luso-Española de Ediciones.

Schank, Roger C. and Robert Abelson (1977). Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum, Hillsdale, NJ.

Schauer, Holger and Udo Hahn (2000). *Phrases as carriers of coherence relations*. In Lila R. Gleitman and Aravind K. Joshi, eds., Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, pp. 429–434. Cognitive Science Society, Lawrence Erlenbaum Associates, Mahwah, New Jersey, Philadelphia, Pennsylvania, USA.

Schiffman, Barry, Inderjeet Mani and Kristian J. Concepcion (2001). *Producing biographical summaries: Combining linguistic knowledge with corpus statistics*. In EACL'01.

Schiffrin, Deborah (1987). Discourse Markers. Cambridge University Press.

Schilder, Frank (2002). *Robust discourse parsing via discourse markers, topicality and position*. In Natural Language Engineering, vol. 8(2&3). Special issue on robust methods in analysis of natural language data.

Scott, Donia R. and Clarisse Sieckenius de Souza (1990). *Getting the message across in RST-based text generation*. In Robert Dale, Chris Mellish and Michael Zock, eds., Current Research in Natural Language Generation, pp. 47–73. Academic Press, New York.

Soricut, Radu and Daniel Marcu (2003). *Sentence level discourse parsing using syntactic and lexical information*. In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL). Edmonton, Canada.

Sparck-Jones, Karen (1993). *What Might Be In A Summary*. Information Retrieval 93: Von der Modellierung zur Anwendung, 9–26.

Sparck-Jones, Karen (1997). *Summarising: Where are we now? where should we go?* In Inderjeet Mani and Mark T. Maybury, eds., Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Assocation for Computational Linguistics. Madrid, Spain.

Sparck-Jones, Karen (1999a). *Automatic summarising: factors and directions*. In Inderjeet Mani and Mark Maybury, eds., Advances in Automatic Text Summarization. MIT Press.

Sparck-Jones, Karen (1999b). *Automatic Summarizing: Factors and Directions*. In Inderjeet Mani and Mark T. Maybury, eds., Advances in Automatic Text Summarization, pp. 1–13. The MIT Press.

Sparck-Jones, Karen (2001a). *Factorial Summary Evaluation*. In Proceedings of the 1st Document Understanding Conference. New Orleans, LA.

Sparck-Jones, Karen (2001b). *Factorial summary evaluation*. In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001. New Orleans, Louisiana.

Spärck-Jones, Karen (2004). *Language and information processing: numbers that count*. evening lecture, ESSLLI 2004.

Stede, Manfred and C. Umbach (1998). *DiMLex: A lexicon of discourse markers for text generation and understanding*. In Proceedings of COLING-ACL '98. Montreal.

Sundaram, H. (2002). Segmentation, Structure Detection and Summarization of Multimedia Sequences. Ph.D. thesis, Graduate School of Arts and Sciences, Columbia University.

Sweetser, E. (1990). From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press.

SweSum (2002). `http://www.nada.kth.se/~xmartin/swesum/index-eng.html`.

Teufel, Simone and Marc Moens (1998). *Sentence extraction and rhetorical classification for flexible abstracts*. In AAAI Spring Symposium on Intelligent Text Summarisation, pp. 16 – 25.

Teufel, Simone and Marc Moens (2002a). *Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status*. In Computational Linguistics, vol. 28(4).

Teufel, Simone and Marc Moens (2002b). *Summarizing scientific articles – experiments with relevance and rhetorical status*. In Computational Linguistics, vol. 28(4). Special Issue on Automatic Summarization.

The R Project for Statistical Computing (2004). *R version 2.0.1*. `http://www.r-project.org/`.

Tucker, Richard (1999). Automatic Summarising and the CLASP system. Ph.D. thesis, University of Cambridge.

Tzoukermann, E., S. Muresan and Judith L. Klavans (2001). *Gist-it: Summarizing email using linguistic knowledge and machine learning*. In ACL-EACL'01 HLT/KM Workshop.

Umbach, Carla (2004). *Contrast and information structure: A focus-based analysis of 'but'*. In Journal of Semantics, vol. 21(2):pp. 155–175.

van Halteren, Hans and Simone Teufel (2003). *Examining the consensus between human summaries: initial experiments with factoid analysis*. In Dragomir R. Radev and Simone Teufel, eds., HLT-NAACL 2003 Workshop: Text Summarization (DUC03), pp. 57–64. Association for Computational Linguistics, Edmonton, Alberta, Canada.

van Halteren, Hans and Simone Teufel (2004). *Evaluating information content by factoid analysis: human annotation and stability*. In EMNLP'04.

Verhagen, Arie (2001). *Subordination and discourse segmentation revisited, or: Why matrix clauses may be more dependent than complements*. In Ted J. M. Sanders, Joost Schilperoord and Wilbert P. M. Spooren, eds., Text Representation. Linguistic and psychological aspects, pp. 337–357. John Benjamins.

Vivaldi, Jorge (2001). Extracción de candidatos a término mediante combinación de estrategias heterog éneas. Ph.D. thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

Vossen, Piek, ed. (1998). Euro WordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers.

Wactlar, H. (2001). *Multi-document summarization and visualization in the informedia digital video library*.

Webber, Bonnie Lynn (1988). *Discourse deixis: Reference to discourse segments*. In Proceedings of the 26th Annual Meeting of the ACL, pp. 113–122. State University of New York at Buffalo.

Webber, Bonnie Lynn, Matthew Stone, Aravind K. Joshi and Alistair Knott (2003). *Anaphora and discourse structure*. In Computational Linguistics.

Witbrock, Michael and Vibhu O. Mittal (1999). *Ultra-summarization: A statistical approach to generating highly condensed nonextractive summaries*. In Proceedings of the

22nd International Conference on Research and Development in Information Retrieval (SIGIR-99).

Yaari, Yaakov (1997). *Segmentation of Expository Texts by Hierarchical Agglomerative Clustering. Technical report also available as cmp-lg/9709015*, Bar-Ilan University, Israel.

Yates, J.A. and W.J. Orlikowski (1993). *Knee-jerk anti-loopism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication. Working Paper 3578-93*, MIT Sloan School. Center for Coordination Science Technical Report 150.

Zajic, David, B. Door and Richard Schwartz (2002). *Automatic headline generation for newspaper stories.* In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization). Philadelphia.

Zechner, Klaus (1997). *A literature survey on information extraction and Text Summarization.* term paper, Carnegie Mellon University.

Zechner, Klaus (2001). Automatic Summarisation of Spoken Dialogues in Unrestricted Domains. Ph.D. thesis, Carnegie Mellon University.