Towards a delimitation of discursive segment for Natural Language Processing applications

Laura Alonso and Irene Castellón

21st December 2001

Abstract

We present an approach to rhetorical segmentation as an aid to automated text summarisation for Spanish. The aim of summarisation is to capture the gist of a text to present it in a reduced form. One of the ways to do so is by extracting the so-called rhetorical structure of the text. Within this framework, we have focused on discourse segmentation in the sentential and subsentential level, a key issue for summarisation subareas like sentence compression or coherence preservation.

Our first task has been delimiting the controversial concept of 'minimal discourse segment'. Dealing with a definite task has enabled us to approach it in a procedural way. Once the working unit has been defined, we have built a tool for automatically segmenting discourse. However concrete as the task may seem, the results and shortcomings of this tool shed light onto key issues for discourse processing, such as the scope of discourse markers, the interaction between syntactical structures, punctuation and cue phrases, etc. Some patterns have been identified that reveal such interactions to be highly informative of discourse structure and yet applicable in the automated processing of some levels of discourse. Moreover, this tool makes it possible to deal with big amounts of data that give a hint of the representativity in real corpora of the phenomena we are dealing with.

1 Introduction

Obtaining the discourse structure of a text has been an area of research for the NLP community since the 1970s ([11, Hobbs 1978], [15, Mann and Thompson 1988]). It has mostly been applied to knowledge-requiring tasks such as natural language generation or machine translation of typologically distant languages. The classical approaches, mainly focused on coherence structure or rhetorical structure, try to obtain a representation of this textual level by heavily relying on expensive sources of artificial knowledge that enable systems to carry out the necessary inferences to derive the rhetorical structure of a text. Thus, applications have just worked for small systems in restricted domains.

However, some approaches to summarisation developed since the mid 1990s supposed a turning point to the way rhetorical structure was exploited in NLP. Traditionally, work in summarisation was divided between knowledge-rich approaches ([24, Schank and Abelson 1977], [18, McKeown and Radev 1995]), working very well but for very restricted domains, or knowledge-poor ones ([14, Luhn 1958]), working poorly but with no restriction as to the texts to handle. The main trend, however, is to merge the advantages of both approaches, so the combination between deep theories and surface techniques has become rather common in working systems.

This mixed approach has made rhetorical structure successfully implementable in a general domain system. Good examples are [19, Ono et al. 1994] and [16, Marcu 1997], who apply the Rhetorical Structure Theory (RST) proposed by [15, Mann and Thompson 1988] by exploiting surface discourse markers so that the rhetorical structure of a text can be obtained. This is done

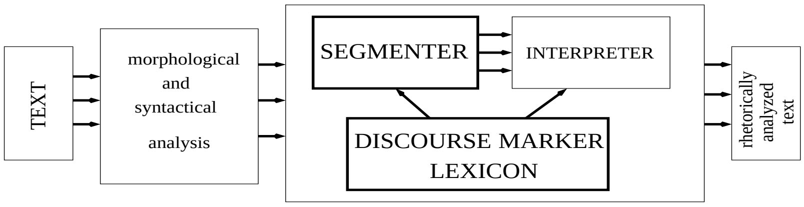

Figure 1: Architecture of the automated rhetorical segmenter

by identifying discourse units and the relationship between them, so that a hierarchical structure can be built that reflects the relative importance of every part of the text. This implies recognising which the most relevant contents are, so that, ultimately, the main communicative intention of the author can be captured.

However, as the authors themselves state, the information provided by discourse markers is often not enough to automatically derive a structure rich enough to build a reliable summary of the text. While surface cues seem to give an accurate account of discourse microstructure[1], they fail to picture a text's macrostructure. Other techniques, such as lexical chains ([4, Barzilay and Elhadad 1997]) or co-reference chains ([3, Azzam et al. 1999]) seem to give a much better account of this discursive level.

Taking this into account, we have focused on obtaining discourse microstructure or local structure for unrestricted Spanish text, relying on surface cues such as cue phrases or syntactical structures. Working in an area where most of the work has been based in English, we have worked specifically for Spanish. We believe that the enriching perspective of linguistic variation can shed new light onto unclear phenomena, such as the interaction between rhetorics and syntax or the role of focus in determining discourse structure. We are currently building a system that can automatically obtain a rhetorical representation of text within a paragraph scope. This system is based in three modules, (see fig 1): the first module determines which the minimal discourse units in a text are, the second one identifies the rhetorical relations holding between them and a lexicon provides the other two with rhetorical lexical information. We have partially implemented two of these, the segmenter and the discourse marker lexicon. Although this tool has been conceived as a module within a larger summarisation system, we believe it can also be useful by itself for other NLP applications.

The rest of the paper is organised as follows. To begin with, we have dealt with determining which the minimal discourse units of a text are. Since this turned out to be too far-reaching an issue for the scope of our work, we constrained the definition of the concept of segment to our particular task, that is to say, relevance assessment and coherence preservation for text summarization. In section 2, we present our approach to discourse segment, discussing some previous work on the topic and trying to achieve a satisfactory delimitation of the concept. We proceed to the presentation of the rhetorical segmentation tool we have developed, based on the concept we have come to. Finally, we discuss the results obtained by the tool and point out some of the issues they reveal.

2 Delimitation of the concept 'minimal rhetorical unit'

Whereas other linguistic levels have been quite satisfactorily accounted for by linguistics and sufficiently formalized for NLP, much work has still to be done to achieve an adequate representation of

[1] We use the terms microstructure and macrostructure in terms of [27, van Dijk 1979].

the discursive dimension of texts. So, it is not surprising that, despite a long tradition on rhetorics and discourse, it is not at all clear what the minimal constituents of a discourse are. From the NLP perspective, many efforts have been done to adequately describe discourse for computational uses. Within this framework, it is commonly assumed that coherent discourse has a compositional nature, being composed by different parts with relations holding between them. However, a common reference point is still missing, as summarised by [5, Carlson and Marcu 2001]:

> "Researchers in the field have proposed a number of competing hypotheses about what constitutes an elementary discourse unit. While some take the elementary units to be **clauses** ([8, Grimes 1975], [7, Givon 1983], [13, Longacre 1983]), others take them to be **prosodic** units ([10, Hirschberg and Litman 1983]), **turns of talk** ([23, Sacks et al. 1974]), **sentences** ([21, Polanyi 1988]), **intentionally** defined discourse segments ([9, Grosz and Sidner 1986]), or the "*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,*" ([22, Polanyi 1996], p.5). Regardless of their theoretical stance, all agree that the elementary discourse units are non-overlapping spans of text."

In general, clauses are considered to be the appropriate size for elementary discourse units. The main shortcoming of clause-centered approaches is that they usually miss a certain discursive aspect of texts in the sub-sentential level. [5, Carlson and Marcu 2001] consider that syntactically smaller units, such as phrases, can be discursive units, although they establish strict conditions for that, for example, that they are signalled by an explicit cue phrase. [26, Schauer and Hahn 2000] address this issue by appealing not to surface clues, but to deep linguistic knowledge. They introduce the idea that rhetorical relations can capture the discursive meaning of those clause constituents which are not in the prototypical argumental structure of a verb.

Seeing that a general, comprehensive definition of segment still seems beyond the possibilities of current research in the field, we have tried to delimit the concept for a concrete task: sentence compression and coherence preservation in aid of automated text summarisation. This meant that we tried to identify those **spans of text that could be removed with no loss in grammaticality or felicity, and whose removal means no significant meaning loss or change**; also those parts of text that could not possibly be removed or separated without causing ill-formedness, or substantial change in discourse. These should be considered minimal discourse units or segments. Being task-oriented, our approach could risk yielding ad-hoc results. We have tried to overcome this both by making our bias explicit and by basing our work in general syntactical and discursive concepts.

It is commonly understood that, to preserve grammaticality, no obligatory argument of a word can be removed. The notion of argumental structure seems to adequately account for what is and what is not an autonomous discourse segment within a clause. As [26, Schauer and Hahn 2000] point out, rhetorical relations may provide an adequate account of those sentence constituents that fall out of the scope of argumental structure. But whereas this may be true for languages such as English or German, where obligatory arguments tend to be realized in the surface representation of a clause, in other languages, such as Spanish or Japanese, the sentential level is not adequate to determine the discursive status of text spans. In this kind of languages it is possible to ellide almost any argument, no matter its function, as long as it is inferrable from context, as can be seen in the following example:

(1) *Apenas recuperado de la fatiga del viaje,* **Juan Pablo II**$_{\mathbf{i}}$ *efectuó ayer un primer balance de* **su peregrinación a Tierra Santa**$_j$. *"Es inenarrable la alegría y el reconocimiento por este don del Señor, tanto tiempo esperado. Ha sido como un regreso a los orígenes, a las raíces de la fe y de la Iglesia", dijo ante millares de personas en la catequesis de la audiencia general de los miércoles en la plaza de San Pedro.* **El Papa**$_{\mathbf{i}}$ ***repasó las etapas de*** **su**

viaje$_j$, *comenzando por la ascensión al monte Nebo, en territorio de Jordania, montaña en la que Moisés avistó la Tierra Prometida antes de morir.*[2]

(1b) **Juan Pablo II**$_i$ *efectuó ayer un primer balance de* **su peregrinación a Tierra Santa**$_j$. **El Papa**$_i$ *repasó las etapas de* **su viaje**$_j$.

(1c) *Juan Pablo II efectuó ayer un primer balance de su peregrinación a Tierra Santa.* ***Repasó las etapas***.

> In (1), the full argument structure of the transitive verb *repasó* is realized. In extracting the most relevant parts of this paragraph (1b), the distance between some constituents of the sentence and their co-referents is reduced, so that, in building a coherent summary (1c), some constituents can be dropped regardless of their function in the sentence, without producing ungrammaticality or meaning change. The most usual case is, like in this example, that the subject is ellided, but this can also affect complements. Also semantically significant components can change their discursive status when text is compressed, like the complement *de su viaje* in the example, which turns from specifying to redundant.

As it follows, it seems that there is a strong rhetorical component in determining what a segment is, with some languages making it more evident than others. Focusing on Spanish, we have tried to determine what is and what is not a distinct segment, and to find those clues that signal discursive status, trying to achieve a compromise between linguistic adequacy and computational implementability. With the notion of discursive removability and autonomy as a starting point, we have tried to identify the linguistic cues that signal discursive status. From those, we have chosen the ones within our NLP tools' processing capabilities, and we have finally designed and partially developed a segmentation tool based on this information.

In order to ensure adequacy to actual linguistic data[3], we carried out an extensive analysis on a corpus of 15 million words, of which 5 million were LEXESP, a representative corpus of written Spanish, and 10 million were newspaper corpus (La Vanguardia and El Periodico). From this corpus, we have studied diverse samples, obtained randomly or oriented to checking hypotheses about discourse structure.

As a result of this analysis, we have found that, in accordance with [9, Grosz and Sidner 1986]'s concept of *satisfaction-precedence*, a span of text is part of the text's argumentative backbone when it provides the necessary context for the following discourse to be correctly interpreted in the framework of the discourse. In any other case, we are in front of a removable discursive segment. The role of syntactic function in discourse structure would then be subordinated to higher discourse principles, either rhetorical, argumentative or both. As for implementability, this can be formally expressed in terms of co-reference chains, scenario setting, necessary event participants, etc. However, obtaining such information requires deep understanding of text, which is far beyond our NLP capabilities. But relying on this delimitation of the concept, we have tried to identify alternative features of segments that we can possibly implement.

[2]*Shortly after recovering from his trip's fatigue, John Paul II made a first balance of his pilgrimage Holy Land. "The joy and acknowledgement are unspeakable for this gift of the Lord, so long awaited. It was like going back to the origins, to the roots of faith and Church", he said before thousands of people in the cathequesis of the wednesday general audience in St. Peter's Square. The Pope revised the stages of his trip, beginning by the ascension to mount Nebo, in Jordanian territory, a mountain where Moses first caught sight of the Promised Land before dying.*

[3]This adequacy should guarantee a good performance of the final application.

We have found that some surface features are good indicators of discursive status. Among those, the ones most useful to our task are cue phrases, since they can provide information not only about segment existence and delimitation, by marking segment boundaries, but also about relations holding between segments. Some syntactical structures, such as relative clauses, non-personal clauses or some subordinate clauses, have proved to signal segments and also simple hypotactic discursive relations. However, we rely on shallow parsing of the input text, which means that many of these structures are not recognised. Alternatively, we rely on some morphological features that are usually combined with these syntactical structures, such as the presence of a relative pronoun. Finally, punctuation marks constitute the major segment boundary indicator in terms of frequency.

However, many of the mentioned clues are highly ambiguous, since they can have a sentential or discursive function. For them to be used in an automated process, it was necessary to determine which instances of each of these features were effectively signalling the presence of a segment. We found that certain combinations of these cues were useful to unambiguously identify them as discourse markers, for example, relative clauses between commas or discursively ambiguous cue phrases immediately preceeded by a punctuation mark. Moreover, the combination procedure led us to consider further discourse markers, which had not been previously considered because they do not serve as such in isolation, such as prepositional or adverbial phrases between punctuation marks. As it follows, simple textual mechanisms, such as punctuation, may play a major role in determining discourse structure, reflecting a richer linguistic information than it has usually been considered.

So, for automated segmentation, a discursive segment should be **any text span within a sentence and no smaller than a phrase marked by any discourse marker or combination of discourse markers that unambiguously signal its status as a removable or autonomous unit**. There are some segments that are not marked as such in surface representation, also, some of the clues we use are highly ambiguous. The scope of certain clues, such as coordinating conjunctions, cannot easily be pre-determined, either. The segmenter, besides its NLP applications, highlighted some of these problematic issues and helped in clarifying the nature and the representativity of some discourse segmentation phenomena through objective data.

3 A segmentation tool

Relying on the delimitation of segment that we have presented in Section 2, we have implemented a segmentation tool designed within a larger system for the automated rhetorical analysis of texts aimed at text summarisation. The objective is identifying minimal discursive units in intrasentential level, that is to say, taking the sentence as the biggest possible unit. Segments are only identified as long as an acceptable level of reliability is maintained. That's why not all possible segments are identified, but only those which are unambigously marked. This strategy prevents errors due to lexical marker ambiguity, however, structural ambiguity still has to be resolved.

We systematised the behaviour of surface clues marking the presence of a segment in two different forms: on the one hand, we built a **cue phrase lexicon** to store lexical discourse markers and segmental information related to them, and, on the other, we developed a **segmentation grammar**. The tool works on a text that has been previously morphosyntactically annotated by the morphological analyzer MACO ([1, Arevalo et al. 2001]), the morphological tagger RELAX ([20, Padro 1997]) and the shallow syntactical analyzer TACAT ([2, Atserias et al. 1998]).

3.1 Architecture of the tool

The general architecture of the tool, as presented in figure 1, consists of two modules. The cue phrase lexicon contains 737 cue phrases that can signal segment boundaries, including conjunc-

discourse marker	boundary	syntactical type	rhetorical type	direction	content
además	not appl.	adverbial	satellizer	inclusion	reinforcement
a pesar de	strong	preposition	satellizer	right	concession
así que	weak	subordinating	chainer	right	consequence
dado que	weak	subordinating	satellizer	right	enablement
o sea	strong	coordinating	connector	bidirectional	reformulation

Table 1: Sample of the cue phrase lexicon

tions, prepositions, adverbs and adverbial phrases. We collected these from three main sources: previous work on automated rhetorical analysis of text ([12, Knott 1996], [17, Marcu 1997]), reference works for Spanish ([28, Martin Zorraquino and Portoles 1999], [1, Arevalo et al. 2001]) and corpus analysis. They are described by six non-exclusive features, of which three can be used by the segmentation module: their function and ambiguity as segment boundaries, their syntactical behaviour and the segment they are more closely related to. Punctuation marks and syntactical structures, which we also consider as discourse markers, are dealt with in the segmentation grammar.

The grammar is a set of 146 context-free rules implemented in the TACAT formalism. They rely on the identification of certain combinations of discourse markers, segments and sentence positions to identify and label segments. We decided to label segments according to their surface form, since, as [6, Corston-Oliver 1999] states, this can help in determining certain rhetorical relations between segments. The working units of the segmentation grammar are:

- **discourse markers**: we use three kinds of items to identify segments: punctuation marks, syntactical structures and cue phrases, stored in a lexicon. This last type of markers can be of two kinds: strong, if they unambiguously mark the presence of a segment, or weak, if their combination with other discourse markers or certain segmental contexts is required to reliably determine the presence of a segment.

- **segments**: we distinguish the following kinds of segments:
  - **plain** text segments: if they are not marked by any discourse marker except full stops. We distinguish paragraph and sentence.
  - **marked** segments: those marked by an explicit discourse marker. We distinguish segments marked by a cue phrase, some syntactical structures and spans of texts between parenthetical signs.

- **sentence positions**: together with punctuation marks, they provide the necessary context to disambiguate weak discourse markers. We distinguish beginning and end positions, where the presence of a comma is sufficient for weak discourse markers to become unambiguous, and middle, where we account for the recursivity of sequences of marked segments and plain text within a sentence. To prevent as many ungrammaticalities as possible, ambiguous clues are only considered to mark a segment in a punctuation context that makes them unambiguous.

An example of how these different types of elements collaborate can be seen in the following set of rules, deriving the analysis at the end. The resulting analysis has been simplified for readability.

4. sentence ==> beginning, comma, middles, end, punt.

8. middles ==> middle, middles.

10. middle ==> text, comma, marked, limit.

14. middle ==> text.

24. limit ==> comma.

43. beginning ==> prepositional.

53. end ==> comma, coordinated.

60. marked ==> relative.

65. coordinated ==> coord, interior.

71. relative ==> relativepronoun, interior.

73. prepositional ==> grup-sp.

29. interior ==> text.

> sentence (**beginning**(**grup-sp**(*Después de tratar de disimularlo*)), **middles**(**middle**(text(*el Presidente tuvo que admitir que el informe*), **marked**(**relative**(**relativepronoun** (*que*) *habían acordado todos los grupos de la oposición*))), **middle**(text(*suponía un revés a su política*))), **end**(**coordinated**(**coord**(*pero*) *reconoció el esfuerzo realizado*)).)[4]

3.2 Results

We carried out two evaluations of our segmentation tool. The first one was purely functional, that is to say, we measured in a small corpus how well the tool did according to its design limitations. We found that 26'73% of segments were not identified at all, and a 4'11% were uncorrectly identified. We used the results of this first evaluation to tune the tool. A second evaluation measured the segmenter's output in comparison to human performance. This second evaluation was carried out on a newspaper corpus of 10613 words, 228 paragraphs and 376 sentences. Text was previously segmented by the tool, and human annotators validated and enhanced this output according to the guidelines defined for the task and their own linguistic intuition.

The corpus was fully annotated by one human judge. Part of the evaluation data was annotated twice, by both the authors. From this partial annotation, preliminary results yield a 87'3% precision and 64'95% recall, (confidence intervals at a 95% significance rate are presented in Table 1). This means that the segmenter identifies less segments than a human in a text, but we are reasonably certain that the segments that are recognised are in accordance to the concept presented in Section 2. This is due to the fact that precision was priorised over recall in designing the tool. The standard F measure obtained is F= 74.5%. [5]

Despite priorising precision, error rates are still quite high, specially for segments that present a very high structural ambiguity, such as coordinated. As for units labelled *weak*, the information they provide can only be considered indicative, since they are too ambiguous in their function as discourse markers. Consequently, error rates in this category should be considered in their appropriate dimension. Nevertheless, the most difficult issue that is still to be solved is structural ambiguity. As it can be seen in example (2), the background problem for solving it is to determine the scope of cue phrases. However, surface cues do not provide the necessary information to determine that unambiguously.

[4] After trying to conceal it, the President had to admit that the report, that all the opposition groups had agreed upon, supposed a setback to his policy, but he recognised the effort carri

[5] The F measure we have used is

$$\left[\frac{2 * precision * recall}{(precision + recall)}\right]$$

type of segment	**precision**	**recall**	number of segments
total	**85’7% - 88’9%**	**62’5% - 67’3%**	1569
adverbial	100%	32’35%	68 (4’3%)
apposition	97’7%	31’03%	87 (5’54%)
coordinated	86’36%	28’18%	110 (7’01%)
weak	82’14%	72’14%	140 (8’92%)
strong	100%	70%	10 (0’64%)
non-personal	87’5%	20%	40 (2’55%)
marked	90’48%	64’29%	42 (2’68%)
prepositional	95’63%	24’59%	183 (11’66%)
relative	96’12%	38’83%	103 (6’56%)
quote	95’65%	94’91%	115 (7’33%)
sentence	92’98%	92’13%	356 (22’69%)
paragraph	93’52%	93’06%	216 (6’94%)

Table 2: Results of the Segmenter

(2a) segmenter’s analysis

*sentence_*[**Los Quince deberían alcanzar un acuerdo político antes que finalice el año ,**

*prepositional_*[según lo pactado en la cumbre de Feira ,
*relative_*[en la que se logró que Luxemburgo aceptara] ,
tras un periodo transitorio de siete años] ,

el intercambio de información sobre los ahorradores ,

*relative_*[lo que se traduce en acabar con el denominado secreto bancario]
.][6]

(2b) human analysis

*sentence_*[**Los Quince deberían alcanzar un acuerdo político antes que finalice el año ,**

*prepositional_*[según lo pactado en la cumbre de Feira ,
*relative_*[en la que se logró que Luxemburgo aceptara ,
*prepositional_*[tras un periodo transitorio de siete años] ,
el intercambio de información sobre los ahorradores ,
*relative_*[lo que se traduce en acabar con el denominado secreto bancario]]] .]

We believe that the segmenter can be significantly improved by incorporating more linguistic knowledge. For example, segment distributional information can still be further exploited, that is to say, that if we enhance the context that we provide in segmentation rules, we will be able to account for more segments than are currently recognised, and with higher reliability. Moreover, we can also exploit segment and discourse marker intrinsic features for disambiguating them: for example, length, lexical weight or rhetorical meaning can help in assessing cue phrase function as segment boundaries. Also morphosyntactical information, such as verb presence, named entities’ segmental patterns, time expressions and other combined clues can contribute to an improved segmentation of discourse.

[6] *The Fifteen should come to a political agreement before the end of the year, as pacted in the Feira summit, where Luxembourg accepted, after a provisional period of seven years, information exchange about bank clients, which means putting an end to the so-called banking secret.*

This results point out some general conflictive issues in discourse structure processing, such as ambiguity in the scope of discourse markers. We have partially solved this by dealing differently with the three different classes of discourse markers that we have distinguished, namely punctuation marks, syntactical structures and cue phrases, and by trying to account for the interactions between them. We have also tried to make a difference between cue phrases according to their rhetorical meaning by distinguishing two main classes: *weak* and *strong*, since they have proved to have very different structural behaviours. However, as the example above shows, this solutions are still insufficient to adequately account for structural ambiguities.

4 Conclusion

We have presented an approach to rhetorical text segmentation for Spanish in aid of automated text summarisation. In the first place, we have tried to define the working unit for sentence compression, a subtask within summarization. This is not a trivial question, since there is no consensus in the field as to what a minimal discourse segment should be. Most of the previous formal studies on discourse have mainly focused on the relationships holding between parts of text, taking clauses as the default unit. However, in contrast with previous computational approaches to discourse, we don't aim at a general representation of discourse, but have a very concrete task. This makes our working unit definiable by procedural means, that is to say, in a bottom-up manner.

Because of this inductive approach and the application's requirement to closely reflect actual linguistic data, we have carried out an extensive corpus study to delimite the concept of discourse segment. Though aiming at particular definition of segment, we have found that general linguistic concepts, such as argument structure, felicity or syntax-discourse interface, are not only useful for adequately describing the phenomena we encounter in corpus, but also necessary to achieve a sound delimitation of the object of study. Finally, the fact that we are working for Spanish introduces a cross-lingual dimension in an area which had traditionally dealt with little linguistic variation. We believe this presents our approach as novel, and that's why it may provide evidence to re-consider some of the aspects that more theoretical approaches to discourse processing take for granted, such as clauses as minimal discourse units or argumental structure as the criterion for defining them.

Based in our delimitation of segment, we have implemented a rhetorical segmentation tool specifically developed for Spanish sentence compression. However concrete the initial target may be, we believe that the information yielded by our system so far can also be useful for a variety of other NLP tasks, such as improving shallow parsers by identifying clause boundaries, providing an improved discourse representation for anaphora resolution systems, determining translation units for machine translation, etc. Preliminary results are promising and already useful for the proposed task. Besides the application point of view, results point out and confirm some conflictive issues in discourse processing: the ambiguity of cue phrases both as to their sentential or discursive usage and as to their rhetorical meaning, structural ambiguity concerning the scope of cue phrases[7] and syntactical structures, the contribution of rhetorical meaning to a hierarchical structure of discourse, and others. The segmentation tool can help in easily quantifying these phenomena by the frequency of the kind of structures in which they occur and the kind of errors they cause. This may suppose a reference for relativising their linguistic dimension.

As future work, we plan to develop a system for interpreting the rhetorical relations between discourse segments. In a first step, we will focus in local discourse structure, trying to improve the cue phrase based method by merging diverse textual information, like morphosyntactical information,

[7]Specially ambiguous as to their scope are those cue phrases having very little rhetorical content, like coordinate conjunctions and commas.

topic-focus patterns, etc. In a second phase, we plan to deal with discourse macrostructure. To obtain an enriched and reliable discourse structure representation in this level, a good approach seems to be the collaboration between different kinds of macrostructural information: co-reference methods (named entity recognition and tracking, lexical chains, anaphora resolution) ([25, Schauer 2000]), thematic structure, high-level rhetoric and argumentative structure, etc. We believe that this integrated approach will not only suppose an improved output of the proposed summarisation system, but will also help in exploring the processing of discourse.

5 Acknowledgements

This research is made possible by grant associated to the X-TRACT project, n º PB98-1226 from the Ministerio de Educación y Cultura of the Spanish Government. Some of the research was conducted as part of the HERMES project, TIC 2000-0335-C032-02. We also want to thank project LEXESP-III, APC 99-0105 of the DGICYT for the availability of a representative corpus of Spanish.

References

[1] Montse Arévalo, Laura Alonso, Mariona Taulé, and M. Antònia Martí. Documentación sobre el analizador morfológico para el castellano (amcas). Technical Report X-Tract 01/01 Working Paper, CLiC (Centre de Llenguatge i Computació), Universitat de Barcelona, Barcelona, 2001.

[2] Jordi Atserias, Irene Castellón, and Montse Civit. Syntactic parsing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada. LREC.

[3] Saliha Azzam, Kevin Humphrey, and Robert Gazauskas. Using coreference chains for text summarisation. University of Maryland, College Park, Maryland, USA, June 1999. ACL.

[4] Regina Barzilay and Michel Elhadad. Using lexical chains for text summarization. In Inderjeet Mani and Mark Maybury, editors, *Intelligent Scalable Text Summarization Workshop (ISTS'97)*, pages 10–17, Madrid, 1997. ACL/EACL.

[5] Lynn Carlson and Daniel Marcu. Discourse tagging manual. Technical Report ISI-TR-545, ISI, 2001.

[6] Simon H. Corston-Oliver. Beyond string matching and cue phrases: Improving the efficiency and coverage in discourse analysis. In *AAAI Spring Symposium on Intelligent Text Summarisation*, pages 9 – 15, 1999.

[7] Talmy Givón, editor. *Topic continuity in discourse: a quantitative cross-language study*. John Benjamins, 1983.

[8] Joseph E. Grimes. *The Thread of Discourse*. Seria Minor. Mouton, 1975.

[9] Barbara Grosz and Candance Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204, 1986.

[10] Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–529, 1993.

[11] J. R. Hobbs. Why is discourse coherent. Technical Report technical note 176, SRI International, Artificial Intelligence Center, 1978.

[12] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Edinburgh, 1996.

[13] Robert E. Longacre. *The Grammar of Discourse*. Plenum Press, 1983.

[14] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159 – 165, 1958.

[15] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organisation. *Text*, 3(8):234–281, 1988.

[16] Daniel Marcu. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 96 – 103, Madrid, Spain, 7 - 12 July 1997.

[17] Daniel Marcu. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 1997.

[18] Kathleen R. McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *ACM Conference on Research and Development in Information Retrieval SIGIR'95*, Seattle, WA.

[19] K. Ono, K. Sumita, and S. Miike. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 344 – 348, Kyoto, Japan, 1994.

[20] L. Padró. *A Hybrid Environment for Syntax-Semantic Tagging*. PhD thesis, Departament de Lenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, 1997.

[21] Livia Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12:601–638, 1988.

[22] Livia Polanyi. The lingustic structure of discourse. Technical Report 96-200, CSLI, 1996.

[23] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696 – 735, 1974.

[24] R. Schank and R. Abelson. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum, Hillsdale, NJ, 1977.

[25] Holger Schauer. Referential structure and coherence structure. In *Proceedings of TALN 2000 — the 7th Conference on Automatic Natural Language Processing*, pages 327–336, Lausanne, Switzerland, October 2000. Association pour le Traitement Automatique des Langues.

[26] Holger Schauer and Udo Hahn. Phrases as carriers of coherence relations. In Lila R. Gleitman and Aravind K. Joshi, editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 429–434, Philadelphia, Pennsylvania, USA, October 2000. Cognitive Science Society, Lawrence Erlenbaum Associates, Mahwah, New Jersey.

[27] Teun A. van Dijk. *Macrostructures: an interdisciplinary study of global structures in discourse, interaction, and cognition*. Erlbaum, Hillsdale, N.J., 1979.

[28] M. Antonia Martín Zorraquino and José Portolés. Los marcadores del discurso. In I. Bosque and V. Demonte, editors, *Gramática Descriptiva de la Lengua Española*, volume III, pages 4051–4213. Espasa Calpe, Madrid, 1999.