An Empirical Approach to Discourse Markers by Clustering

Laura Alonso[1], Irene Castellón[2], Karina Gibert[3], and Lluís Padró[4]

[1] CLiC (Centre de Llenguatge i Computació)
Department of General Linguistics
Universitat de Barcelona
lalonso@lingua.fil.ub.es
[2] Department of General Linguistics
Universitat de Barcelona
castel@lingua.fil.ub.es
[3] Department of Statistics and Operational Research
Universitat Politècnica de Catalunya
karina@eio.upc.es
[4] TALP Research Center
Software Department
Universitat Politècnica de Catalunya
padro@lsi.upc.es

Abstract. The problem of capturing discourse structure for complex NLP tasks has often been addressed by exploiting surface clues that can yield a partial structure of discourse. Discourse Markers (DMs) are among the most popular of these clues because they are both highly informative of discourse structure and have a very low processing cost. However, they present two main problems: first, there is a general lack of consensus about their appropriate characterisation for NLP applications, and secondly, their potential as an unexpensive source of discourse knowledge is weakened by the fact that information associated to them is usually hand-encoded. In this paper we will show how a combination of clustering techniques provides empirical evidence for a characterisation of DMs. This data-driven methodology provides generalisations helpful for reducing the cost of encoding the information associated to DMs, while increasing consistency of their characterisation.

1 Introduction

Some complex Natural Language Processing Applications, such as Machine Translation, Information Extraction, Dialogue Management or Text Summarisation, try to obtain a certain representation of the structure of a text as a whole, what is usually called *discourse*.

Discourse processing tools have traditionally relied on expensive sources of hand-coded knowledge, which implies a high computational cost when applied to real-world tasks. Seeking improvements in efficiency and coverage, some approaches exploit superficial textual clues to obtain a partial representation of discourse.

Cue phrases such as *because*, *although* or *in that case*, usually called Discourse Markers (DMs) are among the most used of these clues. They are highly informative of discourse structure and they can be treated satisfactorily enough with shallow NLP techniques to guide discourse processing tasks such as segmentation, relevance and coherence assessment and even the derivation of a certain structure of discourse [15]. Punctuation, syntactic structures and other shallow textual clues can also have a similar discourse characterising function.

As a drawback to their low processing cost, resources based on DMs are usually built by labour-intensive description and encoding of the information associated to them. In addition, the lack of consensus on the characterisation of DMs has precluded re-usability of these costly resources.

The use of clustering techniques for a description of DMs could be helpful for solving both these problems. First, clustering algorithms can elicit a data-driven organisation of linguistic objects. Secondly, an empirical approach provides non-biased evidence to ground a characterisation of conflictive units such as DMs.

The main goal of clustering techniques is to identify partitions in an unstructured set of objects described by certain characteristics. Those partitions or classes contain similar objects according to some criteria, usually a distance or similarity function. They are expected to be different from each other, altough sometimes they are not, since the method always produces classes, even if they are meaningless. All the objects in a class can be considered together as a whole, and consequently treated in the same way, if the classes can be semantically interpreted by the human analyst.

Therefore, clustering techniques may contribute in characterising DMs by eliciting classes that are empirically grounded and the possible sources of bias that human judgement may introduce. The obtained data organisation is represented in a structure that is easily interpretable by humans. Moreover, it is possible to work with an extensive set of features, comprehending many of the features proposed in heterogeneous approaches, and a high number of examples.

Most of the previous work on obtaining data-driven DM characterisation relies on hand-coded examples [17, 14, 7, 12]. Common to the techniques of clustering and classification based on examples is their capacity of abstracting from a high number of examples and dealing with extensive sets of describing features. The main difference is that classification relies on pre-classified examples, which implies a high cost and unavoidable bias.

The underlying hypothesis of our work is that *DMs with a similar behaviour in naturally occurrent text will correspondingly have a similar behaviour as to the discourse processing instructions they elicit*. As follows, an automated classification of discourse markers according to features describing their occurrences in texts will mirror a taxonomy of the same items as discourse processing devices.

In this paper we present a data-driven classification of DMs in Spanish by clustering techniques. In section 2, we present the data, DMs and corpus, and the clustering tools used. In section 3, we present the results obtained together with a discussion, to finish with conclusions and future work.

2 Experiment

2.1 Discourse Markers

We are working with a set of 577 Spanish DMs, including cue phrases and syntactical structures. These DMs were gathered from previous work about DMs for NLP [15, 13], and specific approaches to DMs in Spanish: grammatical [16], computational (the dictionary of the MACO morphological analyzer for Spanish [3]) and from a corpus study. They are stored in a lexicon of Spanish DMs, with syntactic, discourse segmental and rhetorical information (see Table 1).

DM	boundary	syntactic 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

Table 1. Sample of the cue phrase lexicon

in-lexicon features	**values**
DM form	*aunque, además, así que*, etc.
Rhetorical content	*cause, circumstance*, etc.
Syntactical type	adverb, coordinating, preposition, etc.
Rhetorical type	*connector, satellizer (rhetorically subordinating)*, etc.
contextual features	**values**
Occurrence in initial sentence	yes/no
Occurrence in final sentence	yes/no
Occurrence in initial segment	yes/no
Occurrence in final segment	yes/no
Position of DM in segment	initial, middle, final
Previous word	noun, verb, adverb, adjective, etc.
Following word	grammatical category of the following word
Level of embedding	1 (no embedding), 2, 3, 4, 5, 6
Kind of segment of occurrence	given by the discourse segmenter§
Kind of parent segment	given by the discourse segmenter§
Kind of previous segment	given by the discourse segmenter§
Kind of following segment	given by the discourse segmenter§
Negation in the segment of occurrence	yes/no
Negation in the previous segment	yes/no
Negation in the following segment	yes/no

§ Kinds of segment given by the discourse segmenter: adjectival, adverbial, apposition, unmarked coordinated, marked weak, marked strong, non-personal, cite, marked, parenthetic, prepositional

Table 2. Defining features of DMs for clustering

Each DM instance to be clustered was described by a set of 19 features, upon which the clustering tool evaluated similarity (see Table 2). The choice of features was motivated by previous research on classification of DMs [17, 14, 7], which suggests that discourse structural features (level of embedding, segment markedness, surrounding words, ortography) are useful for describing DM behaviour. We additionally included features productive in the DM lexicon, like syntactical or rhetorical categories, taking care that they did not completely determine the classification.

Another factor that influenced the final choice of the working features was their availability, that is to say, whether they could be easily obtained, from NL Engineering resources such as the DM lexicon mentioned above or by shallow text processing.

2.2 Corpus

We extracted random paragraph-sized occurrences of DMs from a 16 million word corpus (5.5 million words balanced Spanish text (LEXESP) and 10.5 million newspaper text), resulting in a 1,270,993 words corpus with a total of 68,275 instances of DMs. To obtain some of the contextual features listed above, the corpus was previously morphosyntactically analysed [4] and unambiguous intrasentential discursive segments and DMs were identified by an automated discourse segmenter [1].

Unsupervised Corpus All of the used text processing tools priorise precision over recall, which guarantees a high degree of reliability. In the case of DM recognition, however, the segmenter was modified so that all words which were formally identical to a DM were identified, regardless of their ambiguity as to discursive or sentential function, that is to say, insensitive to the fact that a word which was formally identical to a DM might not be performing a DM function in a particular instance.

In order to overcome the capacity limits of the clustering tool Klass+ (see section 2.3), resampling techniques have been used to make bootstrap-oriented clustering. Accordingly, we worked with 5 random samples from this fully automatically annotated corpus, consisting of 200 objects each.

Hand-Tagged Corpus The enhanced recall in DM recognition implied a decrease in precision of approximately 38%. This meant that 38% of the DM instances in the unsupervised samples were performing a non-discursive function in the original text.

To assess the impact of this error rate, we manually tagged a small part of the original corpus, obtaining 277 DM instances with unambiguous discursive function. Two 200-item random samples of this controlled set were clustered. Moreover, classifications of the two hand-tagged samples were taken as comparison ground to better evaluate the adequacy of the describing features.

2.3 Clustering Tools

Among all existing clustering techniques [8], the family of ascendant hierarchical methods (of quadratic cost) is the most popular, since it organizes objects in a binary tree and the number of final clusters may be decided after the clustering. However, other families, like partitioning methods (of linear cost) have become very used, partially because of their capacity to handle huge sets of objects.

The software used to perform the cluster analysis of DMs is KLASS+ [9], an autonomous clustering tool oriented to ill-structured domains. It applies an ascendant hierarchical method [6] that builds classes iteratively clustering the most similar pair of objects at each step[5].

Similarity is calculated according to some distance measure or transformation. KLASS+ permits to work with different distances and similarity coefficients, including mixed distances which enable simultaneously working with categorical and numerical variables. In this application, χ^2 metric was used. It is not order depending, so the final tree is always the same regardless of the objects ordering. In addition, KLASS+ implements *clustering based on rules* method [10] for finding the structure of a dataset.

KLASS+ offers some interpretation-oriented tools for helping in the analysis of the clustering results: using a heuristic criterion, it can recommend the best number of classes, it provides the prototypical description and the distribution of the variables for every class, either in numerical or graphical way, and it can identify *characteristic variables* [11, 10], which can be used to identify the particularities and *meaning* of the final classes.

As an additional aid to overcome the capacity limits of KLASS+, we resorted to Autoclass-C [5], a clustering tool that applies a partitioning method and that could handle the whole set of 68,275 instances of DMs. The outcome of this tool is a non-hierarchical set of clusters and a list of features ordered by their influence values summed over all classes. We used Autoclass-C to further assess the relevance of features in the classes given by KLASS+.

3 Results

First of all, the 2 hand-tagged samples were clustered and a hierarchical tree was found for each of them. Using KLASS+ recommendations, classifications consisting of 3 and 6 classes were obtained. The descriptive tools proposed by KLASS+ showed that negation and segment-contextual features perturbed classification, so they were left out of the objects descriptions and the analysis was repeated without them. After that, we performed classifications of the 5 unsupervised samples. Recommended partitions are still on 3 and 6 classes.

3.1 Variable Relevance

Some of the features in the initial set, like *occurrence in initial/final segment* or *sentence* were not found to be characterising at any level of granularity, since they

[5] Chained reciprocal neighbours is the underlying cluster algorithm.

present very similar distribution in all the classes (Figure 1, left). In contrast, characterising features, like *position of the DM in the segment*, present different values across classes, thus constituting a distinguishing variable of the class, which can be used to identify it either totally or partially (Figure 1, right).

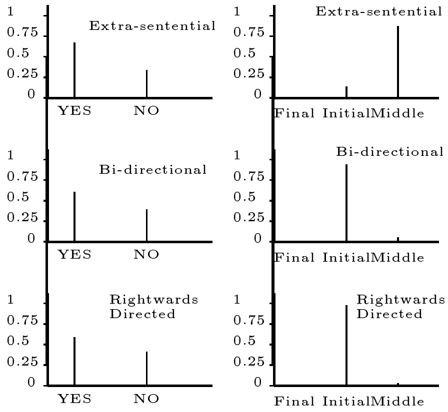

Fig. 1. Distribution across classes for features *occurrence in initial sentence* (left) and *DM position in the segment* (right).

The characterising function of features at different levels of partition motivated a hierarchy of DM classes that can be seen in Figure 2. We compared this hierarchy with the list of feature influence given by Autoclass-C and found a high degree of correlation. As can be seen in Table 3, the main discrepancy between them is that the *type of segment* and *position of DM in the segment* features are considered as more discriminating by Autoclass-C than by KLASS+, whereas *following word* has the opposite interpretation.

The hierarchy is organised as follows:

1. At the topmost level, *position of the DM in the segment* distinguishes DMs occurring mainly in segment initial position from those in other segment positions.
2. DMs integrated in the sentence are further distinguished by *level of embedding*, *type of segment of occurrence* and *type of parent segment*
3. *Syntactical type* subdivides the rightwards directed class, often in correlation with and *rhetorical type*.
4. *rhetorical content* further differentiates classes.
5. At very particular levels, *form of DM* is a highly discriminating feature.

3.2 Consistent Classes

The classifications performed by KLASS+ showed no important differences from one sample to another. Partitions in 3 classes gave the groups:

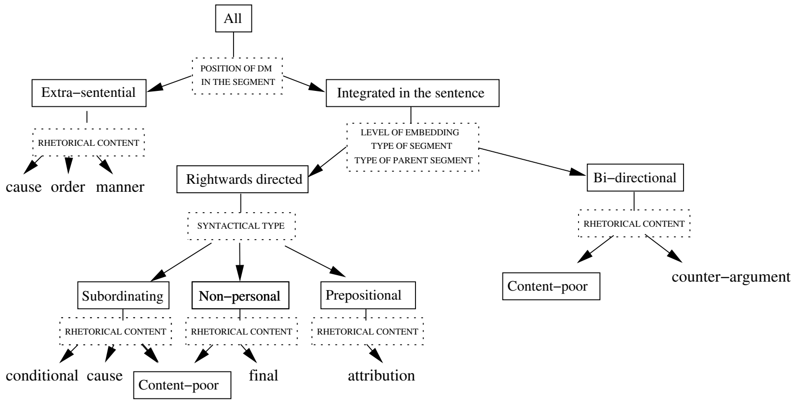

Fig. 2. Hierarchy of features according to their characterising function at subsequent levels of granularity

1. Highly **Bi-directional** DMs, usually in the middle position of a plain text segment. In hand-tagged corpus, relative pronouns are usually clustered with grammaticalized but informative DMs (group 2, below), whereas in unsupervised samples they are classed together with coordinating conjunctions and some instances of very frequent counterargumentative DMs. It seems that this discrepancy succeeds in signalling the non-discursive function of relative pronouns and coordinating conjunctions in the unsupervised sample.
2. **Rightwards directed** DMs, informative of rhetorical structure and content, prototypically found in first or second level of embedding, in initial position of a non-plain text segment. Prepositions, subordinating conjunctions and impersonal forms of verbs are usually classed in this group, although some adverbials can also be found.
3. **Extra-sentential** DMs, carrying strong rhetorical content and often signalling discourse macro-structure, they tend to occur in first or second level of embedding in any kind of segment or segment position. This class is mainly constituted by adverbials and anaphorical expressions.

The 6-class level of generalisation resulted also interesting because feature configurations of groups mirror the granularity of the descriptions of human analysts. In 6-class partitions, we found again the same groups characterised by the core features of the 3-class partitions, but the features *syntactical* and *rhetorical type* of DM made distinctions inside the group of grammaticalized-informative DMs (group 2), as can be seen in Figure 2.

Some other groups found in this level were not characterised by a consistent core of features, and the hierarchy of features was not clear, either. This is a usual phenomenon when clustering ill structured domains [9], like natural language,

KLASS+		Autoclass-C	
relevance	feature	feature	influence
–	position of DM in the segment		
		following word	0.119
-	level of embedding	level of embedding	0.244
-	type of parent segment	type of parent segment	0.248
-	type of segment		
		position of DM in the segment	0.251
+	following word		
+	previous word	previous word	0.313
+	rhetorical type	rhetorical type	0.384
+	syntactical type	syntactical type	0.461
		type of segment	0.469
++	rhetorical content	rhetorical content	0.645
++	form of DM	form of DM	1.000

Table 3. Comparison of feature relevance as given by KLASS+ and Autoclass

since clustering is based on syntactic-like criteria (metrics, similarities) which are not able to capture semantic structures present on data.

Clustering based on rules is a proposal [10] to overcome syntactic limitations by combining the clustering with a partial knowledge base on the domain. The results of the experiment so far enable us to deal with such semantic specifications, the use of rules for finding an improved classification of DMs is currently in progress.

3.3 Supervised *vs.* unsupervised

The main difference between the classifications of unsupervised and hand-tagged samples is that highly ambiguous DMs, such as coordinating conjunctions and relatives, are classed together with extra-sentential DMs in unsupervised classifications, while they are more adequately placed within the grammaticalized DMs in hand-tagged ones. In the latter classifications, a clearcut distinction is made between subordinating and coordinating DMs, both rhetorically and syntactically (Figure 3, right). But in unsupervised classifications, the distinction is made between DMs with vacuous rhetorical content and those having meaningful content (Figure 3, left). The reason for this is that, in hand-tagged samples, these DMs have been correctly characterised in respect to their position in the segment, which is initial when they perform a discursive function. In automated parsing, many might-be-DMs in segment middle position are identified. Consequently, they are classed together with DMs occurring in this position, which are typically extra-sentential.

Taking this into account, hand-tagged classifications can be considered as a reference point or golden standard to assess the discursive function of DMs in unsupervised classifications. As we have shown in [2], divergences between an unsupervised classification and the golden standard allow parametrising prototypicality of DMs. This notion of prototypicality helps in detecting DMs that are possibly misanalysed as to their discursive function.

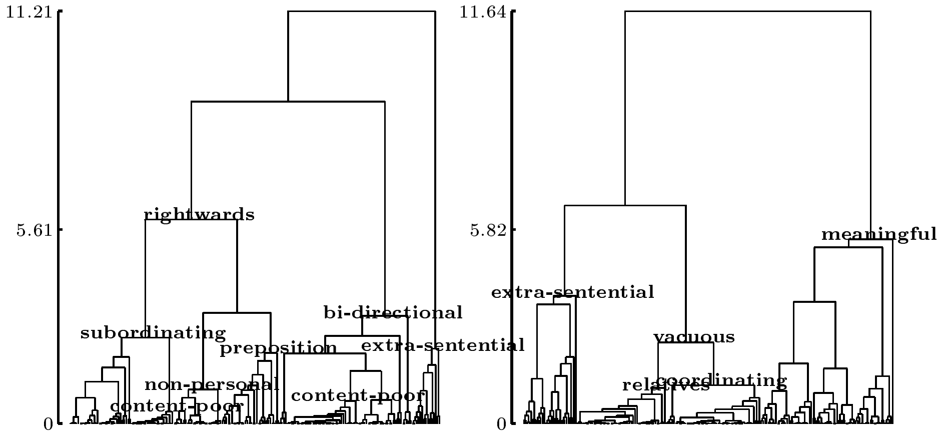

Fig. 3. Right: Dendrogram from a hand-tagged sample, left: Dendrogram from unsupervised corpus

4 Future Work

Once the behaviour of the proposed features has been established, an improved representation of DMs can be pursued. The results of the experiment direct possible enhancements of the set of defining features, by including others that are similar to the ones found more characterising. Moreover, given the success of this preliminary approach to clustering DMs, it seems worthy to perform a deeper analysis of the corpus to obtain DM features that are more informative of discourse processing, such as thematisation, co-reference or an improved representation of polarity.

To overcome capacity limits in clustering, we will:

- Combine the trees of several independent random samples from the same corpus and use the result to estimate the structure of the whole corpus
- Combine complementary clustering tools, for example, by identifying similarity groups within the whole set of DMs with a system capable of dealing with huge amounts of data, like Autoclass-C, and further working them out with a hierarchical tool such as Klass+

The results of clustering will be used to improve the information in the Spanish DMs lexicon. The hierarchical configuration of the features will enable the use of inheritance in a classification of DMs for reducing the cost of encoding information, and it will also guarantee consistency of the data. Another use of clustering is the assessment as to the discursive function of automatically extracted DMs. A typical classification of prototypical DMs (Figure **??**) can be taken as a reference point. To detect DMs not performing a discursive function, a classification of might-be-DMs can be evaluated in comparison to this reference point. Divergences arising between these two classifications should identify conflictive spots where non-discursive elements are most probably located.

5 Conclusions

Our work shows the utility of clustering as a portable and scalable technique for discourse processing technologies. This approach reduces the cost and inconsistency found in extensive use of hand-encoded information, which has already been applied to eliminate redundancies in the DM lexicon presented in Table 1.

The classifications found by KLASS+ have clearly delimited DM groups across a variety of samples. Groups found by clustering correspond to analysts' intuitions, and objects present consistent meanings within classes. Classes are defined by a stable core of features with varying degree of specificity directly related to the granularity of the groups, which can also be expressed in terms of a hierarchy of features. The method of hierarchical clustering has been specially adequate to explore this kind of feature organisation.

Moreover, these groups allow us to determine the degree of prototypicality of DMs. Might-be-DMs with a doubtful discursive function are classed differently in unsupervised and hand-tagged classifications. This has two main implications: First, it constitutes an empirical approach to delimit the concept of DM. Second, hand-tagged classifications can be taken as a comparison ground, so that differences with this reference point serve to signal elements with unprobable discursive function in unsupervised classifications.

6 Acknowledgements

This research has been conducted thanks to a grant associated to the X-TRACT project, PB98-1226 of the Spanish Research Department. It has also been partially funded by projects HERMES (TIC2000-0335-C03-02) and PETRA (TIC2000-1735-C02-02).

References

1. L. Alonso and I. Castellón. Towards a delimitation of discursive segment for natural language processing applications. In *First International Workshop on Semantics, Pragmatics and Rhetoric*, Donostia - San Sebastián, November 2001.
2. L. Alonso, I. Castellón, L. Padró, and K. Gibert. Discourse marker characterisation via clustering: extrapolation from supervised to unsupervised corpora. In *SEPLN*, Valladolid, September 2002.

3. M. Arévalo, L. Alonso, M. Taulé, and M.A. Martí. Documentación sobre el analizador morfológico para el castellano (amcas). Technical Report X-Tract 01/01 Working Paper, CLiC, Universitat de Barcelona, 2001.
4. J. Carmona, S. Cervell, L. Màrquez, M. A. Mart, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An environment for morphosyntactic processing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain, 1998.
5. P. Cheeseman and J. Stutz. Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* AAAI Press/MIT Press, 1996.
6. C. De Rham. La classif. hierarch. selon la méthode des voisins réciproques. *Cahiers d'Analyse des Données*, V(2):135–144, 1997.
7. B. Di Eugenio, J.D. Moore, and M. Paolucci. Learning features that predict cue usage. In *ACL-EACL97, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Madrid, Spain, 1997.
8. B. Everitt. *Cluster Analysis.* Heinemann, London, 1981.
9. K. Gibert. The use of symbolic information in automation of statistical treatment of ill-structured domains. *Artificial Intelligence Communications*, 1997.
10. K. Gibert, T. Aluja, and U. Cortés. Knowledge discovery with clustering based on rules. interpreting results. In *Principles of Data Mining and Knowledge Discovery.* Springer-Verlag, 1998.
11. K. Gibert, U. Cortés, and I. Rodríguez-Roda. Identifying characteristic situations in wastewater treatment plants. In *Workshop in Binding Environmental Sciences and Artificial Intelligence*, 2000.
12. J.H. Kim, M. Glass, and M.W. Evens. Learning use of discourse markers in tutorial dialogue for an intelligent tutoring system. In *COGSCI 2000, Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, Philadelphia, PA, 2000.
13. A. Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations.* PhD thesis, University of Edinburgh, Edinburgh, 1996.
14. D.J. Litman. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94, 1996.
15. D. Marcu. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts.* PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 1997.
16. M.A. Martín Zorraquino and J. Portolés. Los marcadores del discurso. In Ignacio Bosque and Violeta Demonte, editors, *Gramática Descriptiva de la Lengua Española*, volume III, pages 4051–4213. Espasa Calpe, Madrid, 1999.
17. E.V. Siegel and K.R. McKeown. Emergent linguistic rules from inducing decision trees: Disambiguating discourse clue words. In *AAAI94, Proceedings of the 12th Conference of the American Association for Artificial Intelligence*, pages 820–826, 1994.