An Analytic Account of Discourse Markers for Shallow NLP

Laura Alonso, Jennafer Shih, Irene Castellón
Departament de Lingüística General, Universitat de Barcelona
lalonso@lingua.fil.ub.es, shih27@hotmail.com, castel@lingua.fil.ub.es

Lluís Padró
Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
padro@lsi.upc.es

Abstract. We present a feature-based approach to the description of discourse markers (DMs) oriented to automated discourse analysis for shallow Natural Language Processing (NLP). DM describing features have been chosen based on previous work, descriptive adequacy and our concrete NLP needs and capacities. An organization of these features in dimensions of DM meaning has been inferred via data-driven techniques, and finally implemented in a computational DM lexicon.

1 Introduction

Speakers of a language have at their disposal heterogeneous devices that go beyond sentential scope to serve wider communicative purposes: ensuring that communication is established, directing an interaction to achieve certain goals, creating or eliciting relations between the elements in a communicative act (between speakers, a speaker and the message, parts of a message), restricting possible interpretations, etc. Such is the case of certain kinds of pauses and prosodic contours, the choice of syntactic structures or lexic, and also a set of lexical items called "discourse markers" (DMs).

The utility of DMs for systematic, eventually automated, discourse analysis lies in the assumption that they can be consistently associated to a finite set of processing instructions. Thus, they provide rich information at a small processing cost, which makes them a valuable resource for shallow NLP.

However, three main problems problems arise for DMs to be exploited in NLP. First, it is not clear which linguistic units are to be associated with discourse processing instructions, since the concept of DM is hard to define. Then, there is not a consensus set of processing instructions. Finally, the association between DMs and processing instructions is highly ambiguous.

Proceedings of the ESSLLI Workshop The Meaning and Implementation of Discourse Particles
Manfred Stede and Henk Zeevat (editors)

This paper focusses in the two latter problems. The delimitation of a set of processing instructions has been addressed by analyzing the semantics of DMs as a composition of features, which makes the description of DMs more flexible and therefore comparable to diverse sets of processing instructions at various levels of specificity. We have tried to systematise the association between complex processing instructions and DMs by eliciting the organization of semantic features, so that processing instructions can be consistently associated to them.

The rest of the paper is structured as follows: the next Section exposes our working delimitation of DMs. In Section 3, the properties of a feature-based account are discussed, and a set of features describing DM semantics is presented. This proposal is then analyzed via data-driven techniques in Section 4, and an organization of features is established. Section 5 describes how this organization has been implemented in a computational DM lexicon. We finish with some conclusions and future work.

2 Working delimitation of DMs

The concept of "discourse marker" is a controversial one. Therefore, we will not attempt a definition in this paper, we will just expose a working delimitation whereupon our study on DM semantics will build.

A first set of 80 DMs has been compiled. DMs were chosen either because they have been studied in previous work, mostly in Knott (1996) [7] (*however*, *presently*), because they are frequent in corpus (*for example*, *after*), or because they represent a basic discursive meaning (*besides*, *in fact*).

We have only included those DMs for which it was possible to find a near-synonym in English, Spanish and Catalan. The fact that the same or very similar DM can be found **cross-linguistically** is an indication that its discourse meaning is not particular of a language.

Seeking applicability within shallow NLP, only **lexical** items have been considered, because they can be easily identified in a text. They can be either words or multi-words, with functional-semantic unity, allowing discontinuity and morphological variability (e.g., *this is the reason why* and *these are the reasons why*), but leaving apart the discursive effects derived from the semantics of evaluative adjectives and adverbs, the syntax of verbs of knowledge, the effects of aspect and mood contrasts, the distribution of referential expressions and, in general, discourse information provided by open-class lexical items.

In addition, we restricted ourselves to DMs occurring in **written**, **formal** text, mostly of journalistic genre; thus excluding conversational DMs like *aha*. For these DMs, a large number of examples are available, which results very convenient for data-driven analysis.

Finally, DMs that exclusively convey information on relations involving speakers, like *as you know* or *clearly*, have also been excluded. Only DMs conveying **subject-matter** or **argumentative** relations, useful for relevance and coherence assessment, have been considered.

3 A feature-based account of DMs

3.1 Feature-based vs. Relation-based descriptions

DMs have been traditionally described by defining a set of processing instructions or relations and associating each DM to one or more of them. It has often been noted [5, 7] that such a relation-based approach to DM description is problematic, mostly because there is no reference set of discourse relations, and so different descriptions of DMs are usually not comparable.

Knott (1996) [7] analyzes several relation-based theories of text coherence, and develops a set of relations based on the behaviour of DMs in text, which is described with a set of processing primitives. Following this line, we have also taken a data-driven, analytic approach to the description of discourse markers, with two main differences: first, the features proposed here do not appeal to deep understanding of the text, but rely on shallow textual cues, and secondly, we do not aim to obtain a set of discourse relations, but to establish representation of the semantics of DMs.

3.2 DM semantics as a set of features

Features describing the semantics of DMs were defined by the discursive effects of the DM in its context. A set of eight such describing features (Table 3.2) were chosen taking into account:

- previous work on coherence relations, specially from a computational point of view (*causality*, *parallelism*, *revision*, *symmetric*, *asymmetric*).
- application needs, that is to say, discursive information aimed to obtaining a partial structure of discourse that allows for relevance and coherence assessment (*elaboration*, *progression*, *context*, *parallelism*).
- linguistic evidence of the existence of the discursive effect other than the DM signalling it, namely, co-occurrence with other discourse structuring devices, like ortographic correlates of pauses (punctuation marks), modality markers, negation or uneven distributions of lexic or referential expressions.
- objectivity, including only those features for which an acceptable agreement between human judges could be reached, and neutralizing controversial distinctions, like the distinction between *cause* and *consequence*. Tests were established to objectively determine each of the presented discursive effects.

feature	discursive effect	dimension
context	Provides the setting for a discourse entity, as in *figure-ground* [2].	matter
causality	Elicits a causal relation between two elements. It can be underspecified to *context*.	matter
parallelism	Establishes an equivalence between two elements: *A is (mod) B)*, where *mod* is any modifier	matter
revision	Revises an implicature from the previous discourse, proposing a better alternative [7].	matter
elaboration	Does not introduce new information, paraphrases or provides information with no continuity in the discourse, but only restricts possible inferences.	argument
progression	Introduces information that contributes to the progress of the argumentative line.	argument
symmetric	Introduces a textual element at the same level in a hierarchichal structure of discourse.	structure
asymmetric	Introduces a textual element at a different level in that hierarchichal structure.	structure

Table 1.1: DM describing features, defined in terms of discursive effects.

	caus	paral	revision	elab	prog	sym	asym
context	**-0.34**	**-0.60**	-0.02	0.34	-0.33	-0.20	0.18
causality	-	**-0.40**	-0.07	-0.17	0.20	0.07	-0.06
parallelism	-	-	0.12	-0.14	0.10	0.13	-0.12
revision	-	-	-	-0.03	-0.05	0.05	-0.05
elaboration	-	-	-	-	**-0.78**	0.05	-0.05
progression	-	-	-	-	-	0.43	**-0.42**
symmetric	-	-	-	-	-	-	**-0.99**

Table 1.2: Confusion matrix of the correlation coefficient between the distribution of DM semantic features in the description of DMs.

4 Empirical analysis of DM semantics

4.1 Redundancies and correlations: clusters of meaning

Each DM in our working set was assigned positive or negative values for these features according to its core discursive effects in naturally occurring text. The correlation between these features are shown in Table 4.1. Correlation coefficients cose to 0 are not informative of feature organization, but negative correlation values elicit complementary distribution of features, meaning that they belong to the same paradigm. This is clearly the case for *symmetric* and *asymmetric* and for *elaboration* and *progression*, with values close to maximum negative correlation. *Causality*, *context* and *parallelism* can also be thought to belong to the same group.

Taking into account the organization of features elicited by their correlations and also some semantic criteria based on previous work and corpus analysis, features were clustered in three dimensions of DM meaning: **matter**, comparable to [4]'s *internal* or [9]'s *subject-matter*, **argumentation**, describing general mechanisms of discourse planning *à la* Grosz and Sidner [3], and **structure**, that accounts for the hierarchical organization of discourse. The dimension assigned to each feature can be seen in the rightmost column in Table 3.2.

Note that *revision* is the only feature for which a grouping could not be clearly determined, nor by empirical evidence neither by clear semantic criteria. As a matter of fact, this is a very complex feature. Its discursive effects could be analysed as a composition of *elaboration* (the revised element) and *progression* (the newly proposed element). However, we are willing to consider it as a processing primitive because it is usually highly marked, cooccurring with other discourse structuring devices like modality, veritative expressions, etc. Moreover, in many languages there is a very imbricated way of marking revision, evidenced by the wide variety of words signalling this kind of meaning. For example, Yup'ik, a language of the Eskimo-Aleut family, has a concessive mood among a choice of 13 [8].

4.2 Hidden relations: clustering for discovering

To further analyze the behaviour of *revision* and to consolidate the behaviour of the rest of DMs, we resorted to clustering techniques.

Each DM in each of the three languages was described by 13 attributes that have been found characterising of DMs in automated lexical acquisition [1], among which: ratio of occurrences after or between punctuation, at beginning or end of paragraph, length, level of structure, ratio of content words. Values for these attributes were obtained automatically from journalistic corpora, with an average of 1,500 occurrences of each DM.

Different subsets of these attributes were used to build DM-describing vectors. Vectors of DMs in the three languages were clustered with CLUTO[6], so that DMs with most similar values for equivalent attributes were clustered together. To evaluate the descriptive adequacy of different feature subsets, we considered that good clustering solutions should consistently cluster together DMs that had been considered equal in the three languages. To measure this, kappa measures were obtained, recording agreement if equal DMs were clustered together, and disagreement if else. The highest agreement was obtained when DMs were described only by the contextual attributes *ratio of occurrence after punctuation* and *ratio of occurrence before punctuation* (Table 1.1). These attributes have been found to be highly indicative of markerhood in lexical acquisition of DMs as well.

Once the best clustering solution had been determined, the content of each cluster was analyzed with respect to the semantics of the DMs it contained. It was found that DMs conveying revision were consistently clustered

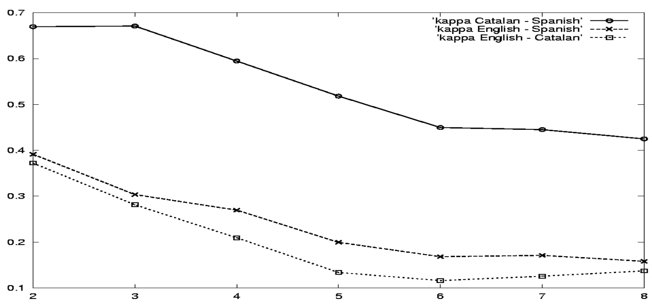

Figure 1.1: Level of agreement (kappa) for cross-lingually equivalent DMs to be clustered together, in clustering solutions where DMs are described by *ratio of occurrence after punctuation* and *ratio of occurrence before punctuation*.

together with those signalling cause, specially asymmetric cause, whereas the rest of meaning combinations are scattered across clusters in a rather asystematic way. This seems to account for a complementary distribution between certain kinds of *cause* and *revision*, which is mirrored by the fact that none of the DMs in the lexicon presents both of these semantic values. This provides enough evidence to consider *revision* within the *matter* dimension of meaning.

5 Implementation of DM semantics in a computational lexicon

Once the organization of DM semantics has been established, it has been implemented in a computational DM lexicon. This resource may serve as a basic knowledge base for a variety of NLP applications working at discourse level. The flexibility of the feature-based description of DMs spares adaptations, since relevant features of DMs can be easily isolated and used for every task. Different levels of analysis are also supported.

One of the main problems in the automated treatment of DMs is their high ambiguity. DMs can be ambiguous with respect to their discursive or sentential function; they can also be polysemous in their discourse semantics. The first kind of ambiguity is addressed by restricting the possible contexts where an ambiguous DM can be assigned a discursive function. These contexts are defined by the same features that were found to be characterising of DMs in lexical acquisition and clustering, namely, contexts with punctuation marks.

Semantic ambiguity of DMs is treated by underspecification at the level of dimensions of meaning. For example, Catalan *perquè* (*because*) is ambiguous with respect to argumentation, expressing either progression or elaboration. Also *by the way* and its correlates in Spanish (*a propósito*) and Catalan (*a propòsit*) are underspecified as to matter and argumentation.

package label	divisions of semantics			DM
	matter	argument	structural	examples
consequence	cause	progression	symmetric	*that's why*
purpose	cause	progression	asymmetric	*in order to*
–	cause	elaboration	symmetric	
reason	cause	elaboration	asymmetric	*due to*
list	parallelism	progression	symmetric	*first*
new argument	parallelism	-	asymmetric	*moreover*
addition	parallelism	elaboration	symmetric	*moreover*
concretion	parallelism	elaboration	asymmetric	*for example*
contrast	revision	progression	symmetric	*but*
counterargument	revision	progression	asymmetric	*however*
concession	revision	elaboration	asymmetric	*although*
–	revision	elaboration	symmetric	
sequence	context	progression	symmetric	*then*
conclusion	context	progression	asymmetric	*in conclusion*
condition	context	elaboration	symmetric	*if*
setting	context	elaboration	asymmetric	*when*

Table 1.3: Relation-like labelling of DM describing features.

Not to sacrifice informativity of such an underspecified description, a set of processing instructions has been associated to types DM, to specify the meaning of a particular DM when there is enough evidence in its context of occurrence. These instructions are associated to the organization of features in dimensions, avoiding *ad hoc* procedures. For example, DMs with with a meaning of parallelism in the matter dimension and with symmetric structural effect (*moreover*) are usually ambiguous with respect to their argumentative effects. It usually depends on their context of occurrence whether they signal *addition*, with effect of elaboration, or a new argument in a *list*, with effect of progression. However, this ambiguity can be successfully solved by taking into account the relevance status of the lexical elements in the considered spans.

Additionally, each combination of features has been associated to a label that resembles relations such as those in widely used theories of discourse coherence, as displayed in Table 5.

This lexicon is currently being implemented in a number of shallow discourse parsing applications. The syntactic and structural features are used for discourse segmentation, to identify discursive segment boundaries. Discourse interpretation for assessment of relevance and coherence, specially oriented to Text Summarization, takes advantage of the matter and argumentation dimensions of DMs. The structural effects of DMs are exploited for deriving a structure of discourse that will be combined with patterns of referential expressions for reference resolution. In all these tasks, rules for specification of DM meaning are applied.

6 Conclusions and Future Work

We have presented a feature-based account of discourse markers oriented to shallow NLP applications. Features have been proposed taking into account previous work, application needs, linguistic evidence and objectivity. The distribution of these features has been analyzed by data-driven techniques, and they have been organized in three dimensions of DM semantics: *argumentation*, *matter* and *structure*.

The proposed description has been implemented in a small computational lexicon in three languages. Usual problems in using DMs in computational applications, like ambiguity, have been adressed by exploiting the organization of the semantic features of DMs in meaning dimensions. This lexicon is currently being used for automated DM acquisition and for shallow discourse analysis, more concretely, to obtain a partial structure of discourse that allows for relevance and coherence assessment.

7 Acknowledgements

This research has been conducted thanks to the grant PB98-1226 of the Spanish Research Department and project PETRA (TIC2000-1735-C02-02).

Bibliography

[1] Laura Alonso, Irene Castellón, and Lluís Padró. X-tractor: A tool for extracting discourse markers. In *LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation*, Las Palmas, 2002.

[2] Nicholas Asher and Alex Lascarides. *The Logic of Conversation*. Cambridge University Press, 2002.

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

[4] M. A. K. Halliday and R. Hasan. *Cohesion in English*. English Language Series. Longman Group Ltd, 1976.

[5] Eduard Hovy and Elisabeth Maier. Parsimonious or profligate: How many and which discourse structure relations? unpublished ms, 1995.

[6] G. Karypis. http://www-users.cs.umn.edu/˜karypis/cluto/index.html, 2002.

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

[8] Steven Lasswell. The alternation of verbal mood in Yup'ik Eskimo narrative. In Marianne Mithun, editor, *Prosody, grammar and discourse in Central Alaskan Yup'ik*, pages 64–97. University of California, Department of Linguistics, Santa Barbara, 1996.

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