# Multiple Sequence Alignment for characterizing the lineal structure of revision

**Laura Alonso**[*], **Irene Castellón**[*], **Jordi Escribano**[†], **Xavier Messeguer**[†], **Lluís Padró**[‡]

[*] GRIAL
Dept. de Lingüística General
Universitat de Barcelona

[†] Software Department
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

[‡] TALP Research Centre
Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

**Abstract**

We present a first approach to the application of a data mining technique, Multiple Sequence Alignment, to the systematization of a polemic aspect of discourse, namely, the expression of contrast, concession, counterargument and semantically similar discursive relations. The representation of the phenomena under study is carried out by very simple techniques, mostly pattern-matching, but the results allow to drive insightful conclusions on the organization of this aspect of discourse: equivalence classes of discourse markers are established, and systematic patterns are discovered, which will be applied in enhancing a discursive parser.

## 1. Motivation

There is practically no consensus on the systematization of the discursive level of language for Natural Language Processing (NLP) applications. The basic discursive concepts (discursive units and discourse structure) are strongly influenced by the standpoint and practical interests of the various approaches. This supposes an important drawback for developing NLP applications that require a certain discursive representation of texts, like Automated Summarization, Dialogue Systems, etc.

In cases where no consensus can be reached, empirical methods can be applied to find theoretically neutral facts in the phenomena under study. These findings can serve as a solid ground whereupon further, deeper studies can be built. The reliability of theoretical claims increases when they are based on neutral facts. Moreover, these facts provide a common, comparable object of study that contributes to the comparability of the claims from different theoretical frameworks.

In this paper we present a first approach to the application of a data mining technique, Multiple Sequence Alignment (MSA), to the systematization of a polemic discourse phenomenon, namely, the expression of *revision*, a family of discourse relations that includes contrast, concession and counterargument. We focus on the study of revision because this phenomenon is very informative of discourse structure, it is highly marked in language and it seems to be mostly explainable as a linear language.

Applying MSA techniques, we establish a methodology for discovering classes of expressions of well-delimited linguistic phenomena, and also certain patterns of behavior. This inferred knowledge is theoretically neutral and can be used to ground theoretical claims or else directly used in NLP applications.

We are willing to work with an amount of data that allows to obtain statistically significant conclusions. This implies working with huge numbers of examples of the phenomena under study. Annotating examples manually, as in Barzilay and Lee, 2002), cannot be done when working with a big number of them. Therefore, examples are analyzed by shallow NLP techniques.

The rest of the paper is structured as follows. In the next Section, the discursive phenomena under study are described, and their linguistic and computational interest is discussed. Section 3. presents our approach to MSA and the tool we have used to carry out the experiments, ALPHAMALIG. In Section 4., we present our procedure for obtaining, representing and mining the data. Then, Section 5. discusses the obtained results, and we finish with some conclusions and future work.

## 2. Discursive strategies: revision

We are going to apply MSA to study a phenomenon in the discursive level of language. Following well-grounded theories of discourse organization, we assume that discursive coherence can be modelled as relations that are established between parts of a text. We are going to focus on one of such coherence relations, what we call *revision*.

We group under the term *revision* a family of discourse relations that share a certain discursive effect, namely, that the propositional or implicational content of one of the related discourse segments is *revised*, usually negated, and the content of the other related discourse segment is proposed as the valid alternative to the revised content:

(1) [$_{seg-1}$ *Although* Greta Garbo was considered the yardstick of beauty,] [$_{seg-2}$ she never married ].
(Lagerwerf, 1998)

In this example, the first segment suggests the expectation that, if a woman is beautiful, she will marry, however, this expectation is negated by the second segment, and both are related by the discourse marker *although*. This kind of phenomena have been widely studied in the literature, under various names: contrast, counterargument, concession, denial of expectation, correction, etc.

### 2.1. The interest of revision for NLP applications

Revision is specially interesting for NLP applications because it provides very insightful information on the structure of discourse, at various levels of analysis: about the structure of discourse, about the argumentative trends, about the relevance of the involved segments, etc. Moreover, revision is less ambiguous than other discourse rela-

This research has been partially funded by the grant PB98-1226 of the Spanish Research Department and by MCyT program - BFF2001-5440.

tions, like for example *cause*, because it tends to co-occur with a wealth of linguistic evidence signalling it.

### 2.2. Revision as a highly marked discursive strategy

Understanding revision in discourse requires costly cognitive processes, because the amount of information and inference mechanisms that are involved in it are very high. As a consequence, revision is highly marked in texts, so as to make it easier for the audience to perform the inference processes intended by the speaker/writer.

The most obvious way of marking revision is by *discourse markers*. Discourse markers are lexical items, with very little variability in their form, that elicit discourse relations between elements in a text. Some examples are *because*, *however*, or *in conclusion*. For example, in the following example, the discourse marker *but* elicits a revision relation between the two discourse segments in the sentence.

(2) [$_{seg-1}$ It is raining today,] [$_{seg-2}$ *but* we are going to the beach anyway ].

However, given that revision involves more costly processes than other discourse relations, various other linguistic devices tend to co-occur with discourse markers in order to clearly signal which of the segments contains the information that is revised, or what subtype of revision relation is intended. Some of these devices are:

**negation** explicitly negating the information to be revised.

(3) [$_{seg-1}$ George Bush is **not** a Nobel Prize holder,] [$_{seg-2}$ *but* a President of the USA].

**modality** placing the information of the segment to be revised in the domain of irreality.

(4) [$_{seg-1}$ This girl **would be** a great researcher,] [$_{seg-2}$ *but* she gets so easily distracted... ]

**evidentiality** questioning the truth status of information in the segment to be revised.

(5) [$_{seg-1}$ **It is true that** the problem is difficult,] [$_{seg-2}$ *but* difficult does not mean impossible].

**quantifiers** restricting the relevant implicatures for a given sentence, and correspondingly restricting the amount of information that may be revised.

(6) [$_{seg-1}$ I enjoy Computational Linguistics,] [$_{seg-2}$ *but* Harry Potter I enjoy **more**].

Some of these linguistic devices are recognizable by simple techniques, like pattern-matching. Therefore, they are very useful in a shallow approach to the representation of revision phenomena, as is our case.

## 3. MSA as a language exploration technique

We argue that MSA, usually applied to DNA sequences, is also useful to study linguistic sequences. Indeed, it has been applied with this aim in a number of cases before: for the discovery of paraphrases for statistical natural language generation (Barzilay and Lee, 2002; Barzilay and Lee, 2003), for the study of word order constraints in different languages (Kruijff, 2002) and to obtain patterns of sentence ordering for the generation of multidocument summaries (Barzilay et al., 2002), among others.

### 3.1. Definition of MSA

MSA is a data mining technique for discovering patterns in a set of comparable sequences. It has been usually applied to DNA sequences, but it can also be used to discover patterns in other kinds of information that can be modelled as a sequence, as is the case of timelines or linguistic production.

The input to MSA are a number of sequences and a similarity criterion or scoring function that describes the similarity between the different symbols that constitute them. Therefore, the modelling of the examples to be studied consists in determining how an example will be translated into a sequence of symbols and the similarity between them.

An alignment algorithm determines the highest-scoring way to perform insertions of gaps, deletions and changes of symbols to obtain a single sequence that subsumes all the input sequences with the least costly changes according to the provided similarity criterion.

One of the reasons why MSA seems well suited for the analysis of linguistic sequences is because it takes into account both the similarity of the sequences under study and their linear configuration, comparable to the semantic and syntactic dimensions of language, respectively.

### 3.2. ALPHAMALIG: a flexible tool for MSA

We have used ALPHAMALIG for aligning examples of revision. In contrast with DNA-oriented tools, Alphamalig supports a **configurable alphabet** and allows determining an explicit, independent **similarity criterion**. It is accessible via web at http://www.lsi.upc.es/~gralggen/recerca/alialfb/alphamalig.html, and provides different possibilities for the visualization of the results. In addition, detailed instructions on usage are available, with examples on the effects of different similarity criteria.

## 4. Mining the data

We obtained 47,000 examples where revision phenomena occurred, from a 6.5 million word journalistic corpus in Spanish. These examples were transformed into alignable sequences by simple techniques, they were grouped in clusters of comparable length, and a similarity criterion was created to stipulate the goodness of match and mismatch between the different elements of the sequences. Since our aim was to study the expression of revision, this similarity criterion was neutral.

Then, sequences were aligned, and the results of the alignment were studied from two perspectives: obtaining patterns from the profile sequences provided by ALPHAMALIG and establishing equivalence classes within the various elements of the sequences, characterized by their contexts of occurrence in alignments, as explained in what follows.

### 4.1. Shallow evidence signalling revision

First, we established a set of shallow cues that signal the presence of revision relations in Spanish text. These cues

were mostly discourse markers, but also linguistic evidence that tends to co-occur with revision, like particles of negation, modality, evidentiality or quantification (see Table 1).

### 4.2. Acquisition and analysis of examples

We identified those sentences in text with the presence of a revision discourse marker, signalling that a revision operation is expressed in that fragment of text. A three-sentence window was extracted for each of the identified sentences, so that each example was constitued by the sentence where the discourse marker was found, the sentence before it and the sentence after it.

Examples were morphosyntactically analyzed with shallow analyzers for Spanish (Atserias et al., 1998), and linguistic elements that are relevant for characterizing revision and detectable by shallow NLP techniques were identified. As can be seen in Figure 1, examples were represented as sequences of these elements, and each of these elements was represented univocally as a letter, which is ALPHAMALIG's expected format, as follows:

- **discourse markers**, each represented by a different symbol (e.g.: *but* $\rightarrow$ *B*)
- **negation particles**, represented by *N*
- **evidentiality particles**, represented by *E*
- **quantifiers**, represented by *Q*
- **verbal phrases**, represented by *V*
- **modal verbs**, represented by *M*
- **punctuation**, commas represented by *C* and periods represented by *P*

### 4.3. Stipulation of the similarity between elements of the sequences

All linguistic elements to characterize examples of revision were represented by a letter, totalling an alphabet of 26 letters. ALPHAMALIG requires that the similarity between the letters of the alphabet is stipulated beforehand. Since our aim was to study the expression of revision, we established a neutral similarity criterion, where each letter in the alphabet had a similarity of 1 with any other symbol.

Additionally, ALPHAMALIG also requires that the similarity of each letter with the gap is stipulated. The gap is the symbol that the tool reserves to represent insertions that are performed in the original sequences for them to match as much as possible with the profile sequence subsuming them all. The similarity of each letter with the gap was set to -10, so that gap insertion was penalized and the rest of elements were forced to match with each other. In this way, the equivalences between symbols were seen more clearly.

Another way of preventing massive insertion of gaps to obtain a profile sequence was aligning only sequences of comparable length. Taking this into account, sequences were grouped according to their length. Most of the sequences had between 10 and 30 elements (9156 sequences between 10 and 15 elements, 13062 between 15 and 20, 16764 between 20 and 30).

### 4.4. Finding equivalence classes of linguistic elements

For each group of sequences aligned with ALPHAMALIG, two kinds of output were obtained: a profile sequence and the set of original sequences, modified so as to be subsumible by the profile sequence.

The profile sequence is a good summary of the most common pattern of elements in the data, so it can be interpreted by itself. The aligned sequences present interesting information about the behavior of the individual elements under study, but the information they present has to be organized for human interpretation.

Each individual element in the aligned sequences was characterized by the *mutual information* it presented with the rest of elements, according to their configuration in the alignment. If the alignment is considered as a matrix, where rows are sequences and columns contain those elements of each sequence that have been considered equivalent for those sequences to match, mutual information is:

$$MI(A,B) = log\frac{P(A,B)}{P(A) * P(B)} \quad (8)$$

where
$n$ = *number of rows*
$m$ = *number of columns*
$pairs = \frac{m*n*(n-1)}{2}$
$P(A,B) = \frac{occurrences\ of\ A\ and\ B\ in\ the\ same\ column}{pairs}$
$P(A) = \frac{occurrences\ of\ A}{m*n}$
$P(B) = \frac{occurrences\ of\ B}{m*n}$

Then, we found equivalence classes of the elements of the sequences by applying clustering techniques. The clustering process was carried out with CLUTO (Karypis, 2002), with the euclidean distance as the similarity function and the *wclink* agglomeration function. In order to be clustered, each element was transformed into a vector, where attributes were the rest of the elements and values for attributes were the mutual information of the given element with each of the other elements in the alphabet.

## 5. Results

### 5.1. Equivalence classes of discourse markers

The presented procedure to establish equivalence classes was only useful for discourse markers and particles of evidentiality, the rest of linguistic evidence seems to be too diverse to be treated homogeneously.

Two main classes of elements were found, differing in the kind of information they were signalling:

**revised information** (*although, in spite of, despite, true that, certainly*)

**well-established information** (*but, however, nevertheless, in fact, in fact$_2$, the fact is that, indeed*)

The rest of elements were not clearly classified. The semantic relation that seems to relate *but$_2$* and *on the contrary* as operators signalling mainly contrast was supported by the fact that they are consistently clustered together, but do not tend to create a cluster of their own, meaning that their behavior is not distinguishable enough or that there are factors that influence their behavior that have not been taken into account in the modelization of the data.

| kind of shallow cue | Spanish | English |
|---|---|---|
| discourse marker | pero, aunque, sin embargo, no obstante, sino, al contrario, a pesar de, con todo, ahora bien, a no ser que, de todos modos, pese a | *but, although, however, nevertheless, but$_2$, on the contrary, in spite of, with all, now, unless, anyway, despite* |
| evidentiality particle | cierto que, en realidad, realmente, verdaderamente, ciertamente, de hecho, el hecho es que, sí | *true that, in fact, actually, really, certainly, in fact$_2$, the fact is that, indeed* |
| negation | no, ningún, ninguno, nada, nadie, sin | *no, no$_2$, none, nothing, nobody, without* |
| quantification | muy, mucho, mucha, muchos, muchas, más, menos, todo, toda, todos, todas, siempre | *very, a lot(inflected), more, less, all (inflected), always* |

Table 1: Shallow cues used to identify and characterize examples of revision relations in text.

(7) *Queríamos ir a la playa . Estaba lloviendo , pero salimos . Al final , el tiempo no estuvo nada malo .*
We wanted to go to the beach . It was raining , but we went out . After all , the weather was not bad at all .
VP . VP , but VP . , neg VP .

**V Z V C B V Z C N V N Z**

Figure 1: Transformation of sentences to sequences of shallow cues relevant for the characterization of revision relations and then to the input format required by ALPHAMALIG.

We believe that the rest of elements suffer from noise in their modelization. While *with all* and *unless* seem to suffer from data sparseness, *now, anyway, actually* and *really* seem to present too diverse behaviors, probably due to their polyfunctionality in language.

### 5.2. Patterns of behavior in the expression of revision

The study of the profile sequences provided by ALPHAMALIG showed that the expression of revision tends to follow a pattern that can be explained as a binary relation between segments, where one segment is presented holding information that will be revised by the other segment. In the most frequent pattern, the first segment tends to contain some or more of the class of particles signalling revised information, as described above. In case more than one of such particles is present, no two discourse markers (*although*) can co-occur, but evidentiality particles tend to co-occur with evidentiality particles (*it is true that*). The second segment tends to present the complementary family of particles, with the same restrictions on co-occurrence. The ordering where the segment presenting the revised information is following the other is less frequent.

No clear patterns of behavior could be found for negation, modal verbs or quantifiers, but we suspect that they significantly contribute to characterize the rest of elements differently.

## 6. Conclusions and Future Work

We have shown how MSA can contribute to systematizing the discursive level of language. This information can be useful for supporting theoretical claims and also for direct application in NLP tools and resources.

In a clearly delimited experiment, we have set a methodology to obtain patterns of behavior and empirically motivated equivalence classes of heterogeneous discourse particles, based on their comparable behavior in alignments. The obtained patterns of behavior will be used for settling and enhancing the scope of a shallow discourse parser, by associating (preferred) subcategorization frames to discourse operators, and also for enhancing the discourse grammar with contextual information.

Future work is aimed at improving the modelization of the examples. We will try incorporate information structure, patterns of pronominalization, etc., inasmuch as they can be treated by shallow NLP techniques.

## 7. References


Atserias, J., J.Carmona, S. Cervell, L. Màrquez, M. A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo, 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *First International Conference on Language Resources and Evaluation (LREC'98)*. Granada, Spain.

Barzilay, Regina and Lillian Lee, 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *EMNLP'02*.

Barzilay, Regina and Lillian Lee, 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *NAACL-HLT*.

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad, 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. In *Journal of Artificial Intelligence Research*.

Karypis, G., 2002. http://www-users.cs.umn.edu/~karypis/cluto/index.html.

Kruijff, Geert-Jan M., 2002. Learning linearization rules from treebanks. invited talk, Formal Grammar'02/COLOGNET-ELSNET Symposium "Combining logical and data-oriented approaches in NLP".

Lagerwerf, Luuk, 1998. *Causal Connectives Have Presuppositions; Effects on Coherence and Discourse Structure*. Den Haag, Holland Academic Graphics.