X-TRACTOR: A Tool For Extracting Discourse Markers

Laura Alonso*, Irene Castellón*, Lluís Padró†

*Department of General Linguistics
Universitat de Barcelona
{lalonso, castel}@lingua.fil.ub.es

†TALP Research Center
Software Department
Universitat Politècnica de Catalunya
padro@lsi.upc.es

Abstract

Discourse Markers (DMs) are among the most popular clues for capturing discourse structure for NLP applications. However, they suffer from inconsistency and uneven coverage. In this paper we present X-TRACTOR, a language-independant system for automatically extracting DMs from plain text. Seeking low processing cost and wide applicability, we have tried to remain independent of any hand-crafted resources, including annotated corpora or NLP tools. Results of an application to Spanish point that this system succeeds in finding new DMs in corpus and ranking them according to their likelihood as DMs. Moreover, due to its modular architecture, X-TRACTOR evidences the specific contribution of each out of a number of parameters to characterise DMs. Therefore, this tool can be used not only for obtaining DM lexicons for heterogeneous purposes, but also for empirically delimiting the concept of DM.

1. Motivation

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 (Marcu, 1997; Dale and Knott, 1995; Kim et al., 2000). Cue phrases such as *because*, *although* or *in that case*, usually called 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 shortcomings: inconsistency in their characterisation and uneven coverage. The lack of consensus about the concept of DM, both theoretically and for NLP applications, is the main cause for these two shortcomings. In this paper, we will show how a knowledge-poor approach to lexical acquisition is useful for addressing both these problems and providing partial solutions to them.

1.1. Delimitation of the concept of DM

A general consensus has not been achieved about the concept of DM. The set of DMs in a language is not delimited, nor by intension neither by extension. But however controversial DM characterisation may be, there is a core of well-defined, prototypical DMs upon which a high consensus can be found in the literature. By studying this lexicon and the behaviour of the lexical units it stores in naturally occurring text, DM characterising features can be discovered. These features can be applied to corpus to obtain lexical items that are similar to the original ones. Applying bootstraping techniques, these newly identified lexical items can be incorporated to the lexicon and this enhanced lexicon can be used for discovering new characterising features. This process can be repeated until the obtained lexical items are not considered valid any more.

It may be argued that enlarging this starting set implies making it more controversial, by adding items whose status as DMs is questionable. However, being empirically grounded, this enlargement is relatively unbiased, and it yields an enhancement of the concept of DM that may be useful for NLP applications.

Taking it to the extreme, unendlessly enhancing the concept of DM implies that anything loosely signalling discourse structure would be considered as a DM. Although this might sound absolutely undesirable, it could be argued that a number of lexical items can be assigned a varying degree of marking strength or *markerhood*[1]. It would be then up to the human expert to determine the load of *markerhood* required for a lexical item to be considered a DM in a determined theoretical framework or application. Lexical acquisition can evidence the load of discursive information in every DM by evaluating it according to the DM characterising features used for extraction.

1.2. Scalability and Portability of DM Resources

Work concerning DMs has been mainly theoretical, and applications to NLP have been mainly oriented to restricted NLGeneration applications. So, DM resources of wide coverage have still to be built. The usual approach to building DM resources is fully manual. For example, DM lexicons are built by gathering and describing DMs from corpus or literature on the subject, a very costly and time-consuming process. Moreover, due to variability among humans, DM lexicons tend to suffer from inconsistency in their extension and intension. To inherent human variability, one must add the general lack of consensus about the appropriate characterisation of DMs for NLP. All this prevents reusability of these costly resources.

[1] By analogy with *termhood*(Kageura and Umino, 1996), which is the term used in terminology extraction to indicate the likelihood that a term candidate is an actual term, we have called *markerhood* the likelihood that a DM candidate is an actual DM.

As a result of the fact that DM resources are built manually, they present uneven coverage of the actual DMs in corpus. More concretely, when working on previously unseen text, it is quite probable that it contains DMs that are not in a manually built DM lexicon. This is a general shortcoming of all knowledge that has to be obtained from corpus, but it becomes more critical with DMs, since they are very sparse in comparison to other kinds of corpus-derived knowledge, such as terminology. As follows, due to the limitations of humans, a lexicon built by mere manual corpus observation will cover a very small number of all possible DMs.

The rest of the paper is organised as follows. In Section 2., we present the architecture of the proposed extraction system, X-TRACTOR, with examples of an application of this system to acquiring a DM lexicon for discourse-based automated text summarisation in Spanish. In Section 2 we present the results obtained for this application, to finish with conclusions and future directions.

2. Proposed Architecture

One of the main aims of this system is to be useful for a variety of tasks or languages. Therefore, we have tried to remain independent of any hand-crafted resources, including annotated texts or NLP tools. Following the line of (Engehard and Pantera, 1994), syntactical information is worked by way of patterns of function words, which are finite and therefore listable. This makes the cost of the system quite low both in terms of processing and human resources.

Focusing on adaptability, the architecture of X-TRACTOR is highly modular. As can be seen in Figure 1, it is based in a language-independent kernel implemented in perl and a number of modules that provide linguistic knowledge.

The input to the system is a starting DM lexicon and a corpus with no linguistic annotation. DM candidates are extracted from corpus by applying linguistic knowledge to it. Two kinds of knowledge can be distinguished: general knowledge from the language and that obtained from a starting DM lexicon.

The DM extraction kernel works in two phases: first, a list of all might-be-DMs in the corpus is obtained, with some characterising features associated to it. A second step consists in ranking DM candidates by their likelihood to be actual markers, or *markerhood*. This ranked list is validated by a human expert, and actual DMs are introduced in the DM lexicon. This enhanced lexicon can be then re-used as input for the system.

In what follows we describe the different parts of X-TRACTOR in detail.

2.1. Linguistic Knowledge

Two kinds of linguistic knowledge are distinguished: general and lexicon-specific. General knowledge is stored in two modules. One of them accounts for the distribution of DMs in naturally occurring text in the form of rules. It is rather language-independant, since it exploits general discursive properties such as the occurrence in discursively salient contexts, like beginning of paragraph or sentence. The second module is a list of stopwords or function words of the language in use.

Lexicon-specific knowledge is obtained from the starting DM lexicon. It also consists of two modules: one containing classes of words that constitute DMs and another with the rules for legally combining these classes of words. We are currently working in an automatic process to induce these rules from the given classes of words and the DMs in the lexicon.

In the application of this system to Spanish, we started with a Spanish DM lexicon consisting of 577 DMs [2]. Since this lexicon is oriented to discourse-based text summarisation, each DM is associated to information useful for the task (see Table 1), such as *rhetoric type*. We adapted the system so that some of this information could also be automatically extracted for the human expert to validate. Results were excellent for the feature of *syntactic type*, and very good for *rhetorical content* and *segment boundary*.

We transformed this lexicon to the kind of knowledge required by X-TRACTOR, and obtained 6 classes of words (adverbs, prepositions, coordinating conjunctions, subordinating conjunctions, pronouns and content words), totalling 603 lexical items, and 102 rules for combining them. For implementation, the words are listed and they are treated by pattern-matching, and the rules are expressed in the form of *if - then - else* conditions on this pattern-matching (see Table 2).

2.2. DM candidate extraction

DM candidates are extracted by applying the above mentioned linguistic knowledge to plain text. Since DMs suffer from data sparseness, it is necessary to work with a huge corpus to obtain a relatively good characterisation of DMs. In the application to Spanish, strings were extracted by at least one of the following conditions:

- Salient location in textual structure: beginning of paragraph, beginning of the sentence, marked by punctuation.
- Words that are typical parts of DMs, such as those having a strong rhetorical content. thetorical content types are similr to those handled in RST (Mann and Thompson, 1988).
- Word patterns, combinations of function words, sometimes also combined with DM-words.

2.3. Assessment of DM-candidate markerood

Once all the possible might-be-DMs are obtained from corpus, they are ponderated as to their *markerhood*, and a ranked list is built.

Different kinds of information are taken into account to assess *markerhood*:

- **Frequency** of occurrence of the DM candidate in corpus, normalised by its length in words and exclusive of stopwords. Normalisation is achieved by the function $normalised\ frequency = length \cdot \log(frequency)$.

[2] We worked with 784 expanded forms corresponding to 577 basic cue phrases

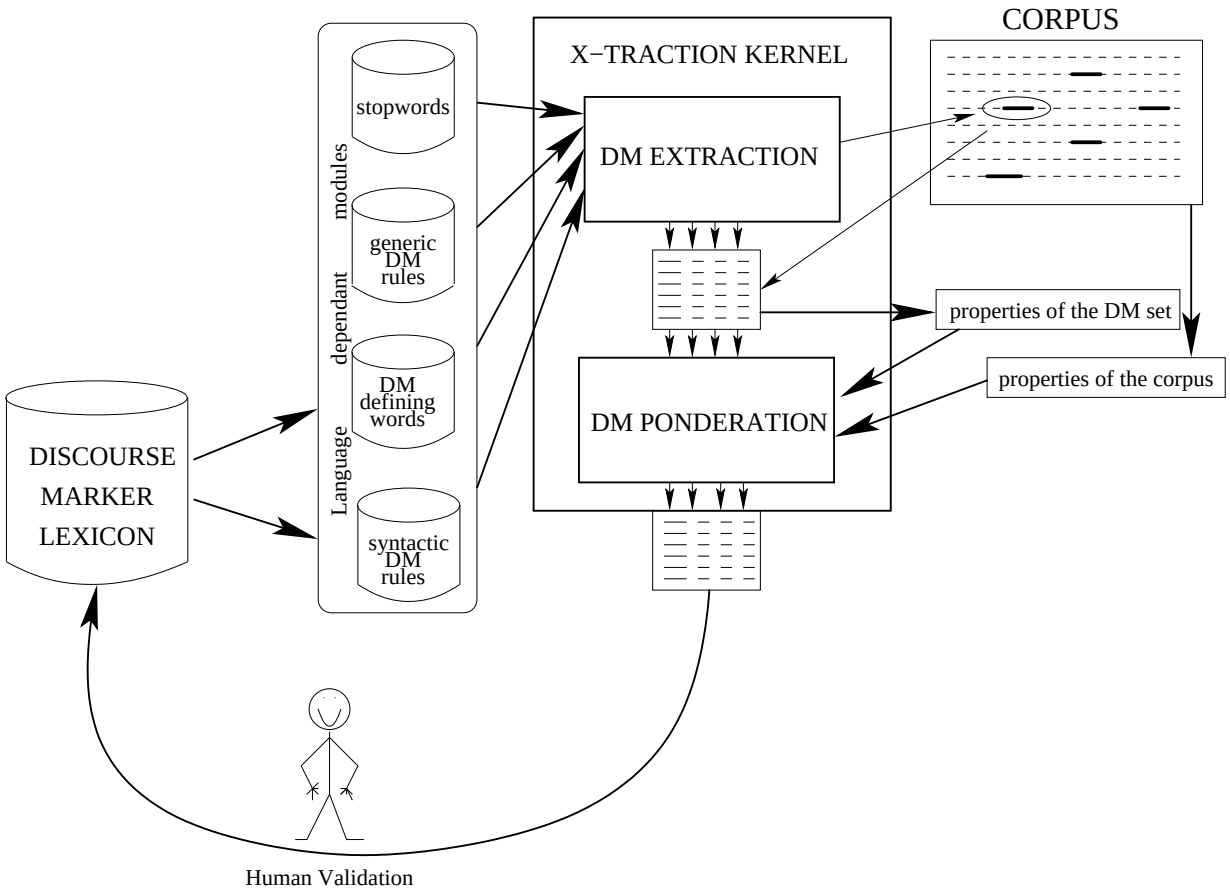

Figure 1: Architecture of X-Tractor

DM	boundary	syntactic type	rhetorical type	direction	con tent
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

- Frequency of occurrence in **discursively salient context**. Discursively salient contexts are preferred occurrence locations for DMs. This parameter has been combined with DM classes motivated by clustering in (Alonso et al., 2002).
- **Mutual Information** of the words forming the DM candidate. Word strings with higher mutual information are supposed to be more plausible lexical units.
- **Internal Structure** of the DM, that is to say, whether it follows one of the rules of combination of DM-words. For this application, X-TRACTOR was aimed at obtaining DMs other than those already in the starting lexicon, therefore, longer well-structured DM candidates were priorised, that is to say, the longer the rule that a DM candidate satisfies, the higher the value of this parameter.
- **Rhetorical Content** of the DM candidate is increased by the number of words with strong rhetorical content it contains. These words are listed in one of the modules of external knowledge, and each has a rhetorical content associated to them. This rhetorical content can be pre-assigned to the DM candidate for the human expert to validate.
- **Lexical Weight** accounts for the the presence of non frequent words in the DM candidate. Unfrequent words make a DM with high *markerhood* more likely as a segment boundary marker.
- **Linking Function** of the DM candidate accounts for its power to link spans of text, mostly by reference.
- **Length** of the DM candidate is relevant for obtaining new DMs if we take into consideration the fact that DMs tend to aggregate.

These parameters are combined by weighted voting for *markerhood* assessment, so that the importance of each of them for the final *markerhood* assessment can be adapted

```
for each word in string
        if word is a preposition, then
          if word-1 is an adverb, then
            if word-2 is a coordinating conjunction, then
              if word+1 is a rhetorical-content word, then
                if word+2 is a preposition, then
                  assign the DM candidate structural weight 5
                elsif word+2 is a subordinating conjunction, then
                  assign the DM candidate structural weight 5
                else assign the DM candidate structural weight 4
              elsif word+1 is a pronoun, then
                assign the DM candidate structural weight 4
            else assign the DM candidate structural weight 3
```

Figure 2: Example of rules for combination of DM-constituing words

to different targets. By assigning a different weight to each one of these parameters, the system can be used for extracting DMs useful for heterogeneous tasks, for example, automated summarisation, anaphora resolution, information extraction, etc.

In the application to Spanish, we were looking for DMs that signal discourse structure useful for automated text summarisation, that is to say, mostly indicators of relevance and coherence relations.

3. Results and Discussion

We ran X-TRACTOR on a sample totalling 350,000 words of Spanish newspaper corpus, and obtained a ranked list of DMs together with information about their syntactical type, rhetorical content and an indication of their potential as segment boundary markers. Only 372 out of the 577 DMs in the DM lexicon could be found in this sample, which indicates that a bigger corpus would provide a better picture of DMs in the language, as will be developed below.

3.1. Evaluation of Results

Evaluation of lexical acquisition systems is a problem still to be solved. Typically, the metrics used are standard IR metrics, namely, *precision* and *recall* of the terms retrieved by an extraction tool evaluated against a document or collection of documents where terms have been identified by human experts (Vivaldi, 2001). Precision accounts for the number of term candidates extracted by the system which have been identified as terms in the corpus, while recall states how many terms in the corpus have been correctly extracted.

This kind of evaluation presents two main problems: first, the bottleneck of hand-tagged data, because a large-scale evaluation implies a costly effort and a long time for manually tagging the evaluation corpus. Secondly, since terms are not well-defined, there is a significant variability between judges, which makes it difficult to evaluate against a sound golden standard.

For the evaluation of DM extraction, these two problems become almost unsolvable. In the first place, DM density in corpus is far lower than term density, which implies that judges should read a huge amount of corpus to identify a number of DMs significant for evaluation. In practical terms, this is almost unaffordable. Moreover, X-TRACTOR's performance is optimised for dealing with huge amounts of corpus. On the other hand, the lack of a reference concept for DM makes inter-judge variability for DM identification even higher than for term identification.

Given these difficulties, we have carried out an alternative evaluation of the presented application of the system. To give a hint of the recall of the obtained DM candidate list, we have found how many of the DMs in the DM lexicon were extracted by X-TRACTOR, and how many of the DM candidates extracted were DMs in the lexicon[3]. To evaluate the goodness of *markerhood* assessment, we have found the ratio of DMs in the lexicon that could be found among the first 100 and 1000 highest ranked DM candidates given by X-TRACTOR. To evaluate the enhancement of the initial set of DMs that was achieved, the 100 highest ranked DMs were manually revised, and we obtained the ratio of actual DMs or strings containing DMs that were not in the DM lexicon. Noise has been calculated as the ratio of non-DMs that can be found among the 100 highest ranked DM candidates.

3.2. Parameter Tuning

To roughly determine which were the parameters more useful for finding the kind of DMs targeted in the presented application, we evaluated the goodness of each single parameter by obtaining the ratio of DMs in the lexicon that could be found within the 100 and 1000 DM candidates ranked highest by that parameter.

In Figure 3 it can be seen that the parameters with best behaviours in isolation are *content*, *structure*, *lexical weight* and *occurrence in pausal context*, although none of them performs above a dummy baseline fed with the same corpus sample. This baseline extracted 1- to 4-word strings after punctuation signs, and ranked them according to their frequency, so that the most frequent were ranked highest. Frequencies of strings were normalised by length, so that $normalised\ frequency = length \cdot \log(frequency)$. Moreover, the frequency of strings containing stopwords was reduced.

[3]We previously checked how many of the DMs in the lexicon could actually be found in corpus, and found that only 386 of them occurred in the 350,000 word sample; this is the upper bound of in-lexicon DM extraction.

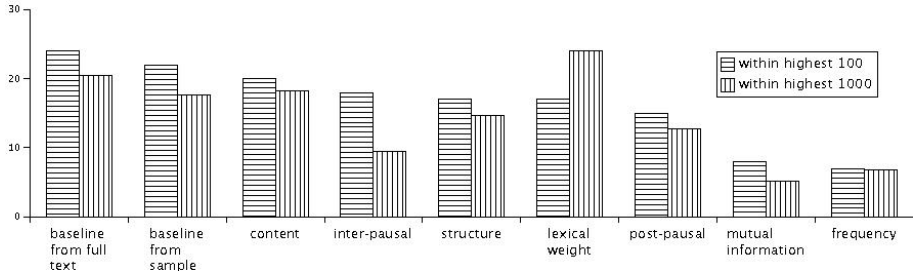

Figure 3: Ratio of DM andidates that contain a DM in the lexicon among the 100 and 1000 highest ranked by each individual parameter

	baseline	X-TRACTOR
Coverage of the DM lexicon	88%	87.5%
ratio of DMs in the lexicon		
within 100 highest ranked	31%	41%
within 1000 highest ranked	21%	21.6%
Noise		
within the 100 highest ranked	57%	32%
Enhancement Ratio		
within the 100 highest ranked	9%	15%

Table 2: Results obtained by X-TRACTOR and the baseline

However, the same dummy baseline performed better when fed with the whole of the newspaper corpus, consisting of 3,5 million words. This, and the bad performance of the parameters that are more dependant on corpus size, like *frequency* and *mutual information*, clearly indicates that the performance of X-TRACTOR, at least for this particular task, will tend to improve when dealing with huge amounts of corpus. This is probably due to the data sparseness that affects DMs.

This evaluation provided a rough intuition of the goodness of each of the parameters, but it failed to capture interactions beteween them. To assess that, we evaluated combinations of parameters by comparing them with the lexicon. We finally came to the conclusion that, for this task, the most useful parameter combination consisted in assigning a very high weight to structural and discourse-contextual information, and a relatively important weight to content and lengh, while no weight at all was assigned to frequency or mutual information. This combination of parameters also provides an empirical approach to the delimitation of the concept of DM, by eliciting the most influential among a set of DM-characterising features.

However, the evaluation of parameters failed to capture the number of DMs non present in the lexicon retrieved by each parameter or combination of parameters. To do that, the highest ranked DM candidates of each of the lists obtained for each parameter or parameter combination should have been revised manually. That's why only the best combinations of parameters were evaluated as to the enhancement of the lexicon they provided.

3.3. Results with combined parameters

In Table 2 the results of the evaluation of X-TRACTOR and the mentioned baseline are presented. From the sample of 350,000 words, the baseline obtained a list of 60,155 DM candidates, while X-TRACTOR proposed 269,824. Obviously, not all of these were actual DMs, but both systems present an 88% coverage of the DMs in the lexicon that are present in this corpus sample, which were 372.

Concerning goodness of DM assessment, it can be seen that 43% of the 100 DM candidates ranked highest by the baseline were or contained actual DMs, while X-TRACTOR achieved a 68%. Out of these, the baseline succeeded in identifying a 9% of DMs that were not in the lexicon, while X-TRACTOR identified a 15%. Moreover, X-TRACTOR identified an 8% of temporal expressions. The fact that they are identified by the same features characterising DMs indicates that they are very likely to be treated in the same way, in spite of heterogeneous discursive content.

In general terms, it can be said that, for this task, X-TRACTOR outperformed the baseline, suceeded in enlarging an initial DM lexicon and obtained quality results and low noise. It seems clear, however, that the dummy baseline is useful for locating DMs in text, although it provides a limited number of them.

4. Conclusions and Future Directions

By this application of X-TRACTOR to a DM extraction task for Spanish, we have shown that bootstrap-based lexical acquisition is a valid method for enhancing a lexicon of DMs, thus improving the limited coverage of the starting resource. The resulting lexicon exploits the properties of the input corpus, so it is highly portable to restricted domains. This high portability can be understood as an equivalent of domain independence.

The use of this empirical methodology circumvents the bias of human judges, and elicits the contribution of a number of parameters to the identification of DMs. Therefore, it can be considered as a data-driven delimitation of the concept of DM. However, the impact of the enhancement obtained by bootstraping the lexicon should be assessed in terms of prototypicality, that is to say, it should be studied how enlarging a starting set of clearly protoypical DMs

may lead to finding less and less prototypical DMs. For an approach to DM prototypicality, see (Alonso et al., 2002).

Future improvements of this tool include applying techinques for interpolation of variables, so that the tuning of the parameters for *markerhood* assessment can be carried out automatically. Also the process of rule induction from the lexicon to the rule module can be automatised, given classes of DM-constituting-words and classes of DMs. Moreover, it has to be evaluated in bigger corpora.

Another line of work consists in exploiting other kinds of knowledge for DM extraction and ponderation. For example, annotated corpora could be used as input, tagged with morphological, syntactical, semantic or even discursive information. The resulting DM candidate list could be pruned by removing proper nouns from it, for example, with the aid of a proper noun data base or *gazetteer* (Arévalo et al., 2002).

To test the portability of the system, it should be applied to other tasks and languages. An experiment to build a DM lexicon for Catalan is currently under progress. To do that, we will try to alternative strategies: one, translating the linguistic knowledge modules to Catalan and directly applying X-TRACTOR to a Catalan corpus, and another, obtaining an initial lexicon by applying the dummy baseline presented here and carrying out the whole bootstrap process.

5. 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).

6. References

Laura Alonso, Irene Castellón, Lluís Padró, and Karina Gibert. 2002. Clustering discourse markers. submitted.

Montse Arévalo, Xavi Carreras, Lluís Màrquez, M.Antònia Martí, Lluís Padró, and M.José Simón. 2002. A proposal for wide-coverage spanish named entity recognition. Technical Report LSI-02-30-R, Dept. LSI, Universitat Politècnica de Catalunya, Barcelona, Spain.

Robert Dale and Alistair Knott. 1995. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

C. Engehard and L. Pantera. 1994. Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27–32.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminolgy*, 3(2):259–289.

Jung Hee Kim, Michael Glass, and Martha W. Evens. 2000. 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.

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

Daniel Marcu. 1997. From discourse structures to text summaries. In Mani and Maybury, editors, *Advances in Automatic Text Summarization*, pages 82 – 88.

Jorge Vivaldi. 2001. *Extracción de candidatos a término mediante combinación de estrategias heterog éneas*. Ph.D. thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.