University of Avignon

From the SelectedWorks of Juan-Manuel Torres-Moreno

Spring 2012

DiSeg 1.0: The first system for Spanish discourse segmentation

Iria da Cunha
Eric SanJuan
Juan-Manuel Torres-Moreno
Marina Lloberes
Irene Castellon

Available at: https://works.bepress.com/juan-manuel_torres-moreno/1/
DiSeg 1.0: The first system for Spanish discourse segmentation

Iria da Cunha a,b,c,e, Eric San Juan b,1, Juan Manuel Torres-Moreno b,c,d,1,2,3, Marina Lloberese e,4, Irene Castellóne,e

1 Institut Universitar de Lingüística Aplicada (Universitat Pompeu Fabra): C/Roc Boronat nº 138, 08018 Barcelona, Spain
2 Universitat de Barcelona: Gran Via de les Corts Catalanes nº 585, 08007 Barcelona, Spain
3 École Polytechnique de Montréal: C.P. 6079, succ. Centre-ville, Montréal, Québec, Canada H3C 3A7
4 Laboratoire Informatique d’Avignon (Université d’Avignon et des Pays de Vaucluse): 339, chemin des Meinajaries, Agroparc, BP 91228, 84911 Avignon, Cedex 9, France
5 Instituto de Ingeniería (Universidad Nacional Autónoma de México): Circuito Escolar sn, Ciudad Universitaria, CP 04510, Delegación Coyoacán, México D.F., Mexico

E-mail addresses: iria.dacunha@upf.edu (I. da Cunha), eric.sanjuan@univ-avignon.fr (E. San Juan), juan-manuel.torres@univ-avignon.fr (J.M. Torres-Moreno), marina.lloberese@gmail.com (M. Lloberese), icastellon@pub.edu (I. Castellóne).

1 Tel.: +34 935 422 322; fax: +34 935 422 321.
2 Tel.: +1 514 340 4711; fax: +1 514 340 5859.
3 Tel.: +34 93 403 4695; fax: +34 93 318 9822.
4 Tel.: +33 4 90 84 35 09; fax: +33 4 90 84 35 01.
5 Tel.: +33 4 90 84 35 09; fax: +33 4 90 84 35 01.

A R T I C L E   I N F O

Keywords: Discourse parsing Discourse segmentation Shallow parsing Rhetorical Structure Theory

A B S T R A C T

Nowadays discourse parsing is a very prominent research topic. However, there is not a discourse parser for Spanish texts. The first stage in order to develop this tool is discourse segmentation. In this work, we present DiSeg, the first discourse segmenter for Spanish, which uses the framework of Rhetorical Structure Theory and is based on lexical and syntactic rules. We describe the system and we evaluate its performance against a gold standard corpus, divided in a medical and a terminological subcorpus. We obtain promising results, which means that discourse segmentation is possible using shallow parsing.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays discourse parsing is a very prominent research topic, since it is useful for text generation, automatic summarization, automatic translation, information extraction, etc. There are several discourse parsers for English (Marcu, 2000a, 2000b), Japanese (Sumita, Ono, Chino, Ukita, & Amano, 1992) and Brazilian Portuguese (Pardo & Nunes, 2008; Pardo, Nunes, & Rino, 2004). Most of them use the framework of Rhetorical Structure Theory (RST) (Mann & Thompson, 1988). However, there is no discourse parser for Spanish. The first stage in order to develop this tool for this language is to carry out discourse segmentation automatically. As stated in Tofiloski, Brooke, and Taboada (2009): “Discourse segmentation is the process of decomposing discourse into Elementary Discourse Units (EDUs), which may be simple sentences or clauses in a complex sentence, and from which discourse trees are constructed”. There are discourse segmenters for English (Soricut & Marcu, 2003; Tofiloski et al., 2009)6, Brazilian Portuguese (Mazeiro, Pardo, & Nunes, 2007)7 and French (Afantenos, Denis, Muller, & Danlos, 2010), but not for Spanish. All of them require some syntactic analysis of the sentences. The segmenters developed by Soricut and Marcu (2003), Mazeiro et al. (2007) and Tofiloski et al. (2009) rely on a set of linguistic rules. The one by Afantenos et al. (2010) relies on machine learning techniques: it learns rules automatically from thoroughly annotated texts. This system obtains good results and its authors use a very interesting strategy. Nevertheless, this type of approach (with machine learning) needs a very high quantity of texts to be able to “learn” and to perform a suitable segmentation. Segmenters using symbolic strategies (with syntactic and lexical rules) could give material to these systems. It is difficult to evaluate the results obtained by a segmenter based on machine learning, because it can be applied over millions of texts. It is easier to evaluate the segmentation quality of symbolic segmenters and, later, to use this segmented corpus as a learning corpus in order to develop a discourse segmenter. In this way, a combined strategy (linguistic and statistical) could be employed.

In this paper we present DiSeg, the first discourse segmenter for Spanish. It produces state of the art results while it does not require syntactic analysis but only shallow parsing with a reduced set of linguistic rules. Therefore it can be easily included in applications requiring fast text analysis on the fly. In particular it will be part of the discourse parser for Spanish that we are carrying out. It will be also used in tasks involving human discourse annotation, since it will allow annotators to perform their analysis starting from a unique automatic segmentation. We describe the system, based on shallow parsing and syntactic rules that insert...
segment boundaries into the sentences. Like Afantenos et al. (2010) and Tofiloski et al. (2009), we evaluate the system performance over a corpus of manually annotated texts. We obtain promising results, although the system should be improved in some aspects.

The rest of the paper is structured as follows. Section 2 explains our methodology. Section 3 describes the grammar used. Section 4 details the implementation. The gold standard corpus and the results are explained in Section 5. In Section 6 we include the discussion about the system limitations, with some examples. Finally, Section 7 is devoted to conclusions and future work.

2. Materials and methods

The theoretical framework of our research is based on Rhetorical Structure Theory (RST), as defined by Mann and Thompson (1988). As mentioned in Taboada and Mann (2005), this theory is used in a wide range of Natural Language Processing (NLP) applications like automatic text generation (Dale, Hovy, Rösnér, & Stock, 1992; Hovy, 1993; O’Donnell, Mellish, Oberlander, & Knott, 2001), summarization (Marcu, 2000a, 2000b; Pardo & Rino, 2002; Radev, 2000), translation (Ghorbel, Ballim, & Coray, 2001; Marcu, Carlson, & Watanabe, 2000), etc. In all these applications, RST is used to obtain a deeper linguistic analysis. Fig. 1 shows an example of RST discourse tree.

As it can be seen in this example, RST can work at two levels. At sentence level to analyze them, or at an upper level to relate them. We shall not consider the transversal case where sub-units inside a sentence can be individually related to units inside other sentences. In da Cunha et al. (2007) we have shown how RST at upper level can improve automatic summarization methods based on full sentence selection from the source text by scoring them according to their discursive role. In that paper we focus on RST applied at sentence level.

RST sentence segmentation tools are necessary for further discursive analysis but they are also useful on their own in many NLP applications. For example, by segmenting complex sentences they can be used in sentence compression. Therefore, in automatic summarization, RST-based strategies would allow to eliminate some passages of these sentences, obtaining more suitable summaries. With regard to automatic translation, most usual strategies rely on statistical sentence alignment. Again, for complex sentences, the results of these statistical systems could be improved by aligning sub-discourse units. Moreover, fast segmentation tools based on shallow parsing, as the one we propose here, can be applied in focus Information Retrieval systems that have to return short text passages instead of complete documents.

Let us now precise the notion of EDU in our work. We consider them as in Carlson and Marcu (2001), but with some differences, similar to those included in da Cunha and Iruskieta (2010) and Tofiloski et al. (2009). The aim of these differences is to be able to clearly differentiate syntactic and discursive levels. In this work, we consider that EDUs must include at least one verb (that is, they have to constitute a sentence or a clause) and must show, strictly speaking, a rhetorical relation (many times marked with a discourse connector).

For example, sentence 1a would be separated into two EDUs, while sentence 1b would constitute a single EDU:

1a [The hospital is adequate to adults,]_EDU1_ [but children can use it as well,]_EDU2_
1b [The hospital is adequate to adults, as well to children,]_EDU1_

Furthermore, subject and object clauses are not necessarily considered as EDUs. For example, sentence 2 would be a single EDU:

2 [She indicated that the emergency services of this hospital were very efficient,]_EDU1_

We have then developed a segmentation tool based on a set of discourse segmentation rules using lexical and syntactic features. These rules are based on:

− discourse markers, as “while” (mientras que), “although” (aunque) or “that is” (es decir), which usually mark relations of Contrast, Concession and Reformulation, respectively (we use the set of discourse markers listed in Alonso (2005);
− conjunctions, as, for example, “and” (y) or “but” (pero);
− adverbs, as “anyway” (de todas maneras);
− verbal forms, as gerunds, finite verbs, etc.;
− punctuation marks, as parenthesis or dashes.

Finally, we have also annotated manually a corpus of texts to be used as gold standard for evaluation. The elaboration of a gold standard was necessary due to the current lack of discourse segmenters for Spanish. We thus evaluate DiSeg performance, measuring precision, recall and F-Score over this annotated corpus. We also consider three different baseline systems and a simplified system named DiSeg-base.

3. Calculation

In this section we present the grammar we have used to develop our discourse segmenter, and we show how its implementation has been carried out.

3.1. Grammar

Discourse markers categorization task in DiSeg relies on the chunker (Castellón, Civit, & Aserias, 1998; Civit, 2003) developed as a resource for the FreeLing environment, an open-source multilingual NLP library (Aserias et al., 2006; Padró, Collado, Reese, Lloberes, & Castellón, 2010). The FreeLing chunker is a bottom-up parser based on Hidden Markov Model (HMM) classifiers, which uses a Context Free Grammar (CFG) and provides shallow parsed trees or chunks with relative depth. The implemented algorithm assigns a priority to each chunk. It determines that the longest and deepest chunks are applied first (Fig. 2). The FreeLing chunking

---

Fig. 1. An example of RST discourse tree. It contains three Nucleus–Satellite relations (cause, antithesis and reformulation) and one multinuclear relation (list).
grammar includes a set of options to perform extra tasks over parse trees, such as hiding intermediate categories in recursive trees, adding shallowness to relative deep trees, etc. Chunking rules have been designed in order to work with terminal and pre-terminal nodes, syntactic categories, lemmata and word forms.

Some new rules (41 rules) have been added to the FreeLing chunking grammar in order to recognize some expressions which are considered discourse markers candidates. Discourse markers are classified into two groups: non-ambiguous and ambiguous. Both classes include single word forms (ex.: concretamente “specif-
ically”, también “also”), phrases (ex.: en resumen “to summarize”, por ejemplo “for example”), and composed conjunctions (ex.: asi como tambi¢n “also”).

The implementation of all this information in the grammar is carried out by specifying external lexicons or assigning categories to chunks.

disc-mk → RG(‘‘lista \_sadv.txt’’). [1]
disc-mk → +SPS00(en). NCFS000(realidad). [2]

The rule [1] has two constraints that the expression analyzed by the chunker must meet. Firstly, this expression must be an adverb (RG). Secondly, this expression must be included in the adverbial expressions lexicon (lista\_adv.txt).

Example [2] shows a rule used in chunk categorization. This rule defines prepositional phrase parsing between the preposition en (“in”) and the noun realidad (“reality”) (en realidad “in fact”). As shown in this example, the parenthesis operator ‘‘(’’ means that words within parenthesis must be word forms instead of lemmas.

Regarding ambiguous discourse markers (ex.: por lo cual “as a consequence”), the grammar only detects them, but they can only be disambiguated in the next task by means of rules taking into account the context.

disc-mk-amb → +SPS00(por). DA0NS0(lo).PROCNO000(que). (3)

In [3], the rule illustrated recognizes the expression por lo que (“as a consequence”) as a chunk that may be an ambiguous discourse marker. In other words, in order to solve the categorization of this expression, the context of the sentence needs to be checked.

3.2. Implementation

As we have explained, DiSeg implementation relies on Freeling, although we have carried out some modifications into the default grammar of the shallow parser (mainly recategorizations of some elements into discourse markers). Freeling output is then encoded into an XML structure to be processed by perl programs that apply the discourse segmentation rules in a two-step process.

First (DiSeg-base), candidate segment boundaries are detected using two simple automata based on the following tags: ger, forma\_ger, ger\_pas (that is, all possible present participles or gerunds), verb (that is, finite verbal forms), coord (coordinating conjunctions), conj\_subord (subordinating conjunctions), disc\_mk (recategorized elements) and grup\_sp\_inf (infinitives). The only text markers that are used apart from these tags are the period and two words: que (“that”) and para (“for”).

Second (DiSeg), EDUs are defined using a reverse parsing from right to left where boundaries are considered only if there is a verb in the resulting segments before and after this boundary. Indeed, if all previously inserted boundaries were considered, EDUs without verbs could be generated. Thus, the architecture of the system has several stages:

SENTENCE SEGMENTATION (with FreeLing)
↓
SHALLOW PARSING (with a recategorized FreeLing grammar)
↓
TRANSFORMATION TO XML (with perl programs)
↓
SEGMENTATION RULES APPLICATION (with perl and twig)

DETECTION OF SEGMENT BOUNDARIES/EDUS DEFINITION

Fig. 3 shows a screenshot of DiSeg-1.0 system. It requires Freeling and it is made of three elements:
1. A grammar for Freeling.
2. A small perl program to transform FreeLing output into XML.
3. A second perl program that applies the discourse segmentation rules and requires TWIG library for XML.

Appendix A includes a passage of the gold standard corpus, its translation and its DiSeg segmentation. Appendix B shows the resulting XML source with:

- Tags inserted by FreeLing, including disc\_mk tags resulting from the added rules to the default FreeLing CFG for Spanish.
- The segD tags corresponding to possible EDU boundaries (DiSeg-base).
- The subseg tags corresponding to selected EDUs (DiSeg).

---

Fig. 2. FreeLing chunking output for the sentence Se distribuyeron periódicos a pesar de los piquetes (“Newspapers are distributed despite pickets”).
The XML elements `<seg rule="rule">...<seg>` that also appear in this output indicate the passages that DiSeg had to analyze. The analysis is carried out inside these elements. In the current case, these elements are full sentences but, whenever punctuation is ambiguous, extended elements could be considered.

Since we use very few text marks, our approach should be easily adapted to other Latin languages defined in FreeLing. Moreover, DiSeg-base could be implemented in a CFG, but it would be less computationally efficient. It is only the final reverse parsing that is not CFG definable. In our experiments we have tested to what extent the non CFG module is necessary.

### 4. Results

In this section we present the gold standard test corpus that we have compiled to perform the evaluation of the system; moreover we show the obtained results, over the medical and terminological sub-corpora.

#### 4.1. Gold standard

The gold standard test corpus includes two sub-corpora. The first one consists of 20 human annotated abstracts of medical research articles. These abstracts were extracted from the on-line *Gaceta Médica de Bilbao* ("Medical Journal of Bilbao"). As Table 1 shows, this sub-corpus includes 169 sentences, 3981 words and 203 EDUs. The statistics included into this table are similar to the statistics of the gold standard corpus used by Tofiloski et al. (2009) to develop a discourse segmenter for English, obtaining very good results. This sub-corpus was segmented by one of the authors of this paper (following the guidelines of our project). Another linguist, external to the project, segmented the corpus following the same guidelines. We calculated the precision and recall of this second annotation. Both measures were very high: precision was 98.05 and recall 99.03. Moreover, after short discussions between annota-

---

*Fig. 3.* A screenshot of DiSeg-1.0 system. The system can be used on-line at [http://www.daniel.iut.univ-metz.fr/DiSeg/WebDiSeg/](http://www.daniel.iut.univ-metz.fr/DiSeg/WebDiSeg/) and is also available under General Public License (GPL).
We also consider a simplified system named DiSeg-base, where all candidate boundaries are considered as real EDU ones, even though some generated segments can have no verbs.

For Baseline_1, Baseline_2 and DiSeg-base we do not count sentence boundaries.

Our results show that DiSeg full system outperforms DiSeg-base and all the baselines. DiSeg obtains an F-Score of 80% (71% of precision and 98% of recall), while DiSeg-base obtains an F-Score of 74% (70% of precision and 80% of recall). The results obtained by the three baselines are lower: 72% of F-score by Baseline_2 (68% of precision and 82% of recall), 39% of F-score by Baseline_1 (33% of precision and 70% of recall) and 62% of F-score by Baseline_0 (100% of precision and 49% of recall).

F-Score differences are statistically significant according to the pairwise Student test at 0.05 between the two versions of DiSeg and at 0.01 among DiSeg and the three baselines (Baseline_2, the most sophisticated baseline, obtains the best results).

These results are similar to those obtained by the discourse segmenter for English developed by Tofiloski et al. (2009): 93% of precision, 74% of recall and 83% of F-Score.

Although we considered these quantitative results were good, we carried out a qualitative analysis in order to detect the main performance problems, which we detail in Section 5. After this qualitative analysis of the results, we developed three more symbolic rules to try to solve the main systematic segmentation errors (see Section 5 for details). The rules are applied in the post-processing stage (EDUs definition), so they can increase results precision. In this way we try to optimize the system, that we call now DiSeg-1.0. We have applied DiSeg-1.0 over the same medical sub-corpus, in order to check if the results improve. The results of DiSeg-1.0 outperform the results of the previous version of the system, DiSeg. The F-score of DiSeg 1.0 is 96% (97% of precision and 96 of recall) and the F-score of DiSeg is 80%, so there is a difference of 16 points.

### 4.3. Experiments with the terminological sub-corpus

Once we have tested the system with the medical sub-corpus, we have decided to apply it over another corpus including documents from a very different domain, the terminological one. Both sub-corpora are specialized, but they correspond to very different domains: a technical or scientific domain vs. a humanistic or linguistic domain. The reason of this selection was that we wanted to check that DiSeg-1.0 is suitable to segment any kind of text.

We have applied DiSeg-1.0 over this terminological sub-corpus. We have used a baseline to carry out this evaluation. This Baseline_0 was performed in the same way we have done in the previous experiment with the medical sub-corpus. In this experiment, DiSeg-1.0 obtains an F-score of 91% (95% of precision and 87% of recall) and Baseline_0 obtains an F-score of 62% (100% of precision and 49% of recall). Thus, the results obtained by our system are very high (much more than the baseline), and this means that our system can be used to segment texts of different domains.

### 5. Discussion

As we have mention in Section 4, after the first experiment with the medical sub-corpus, we carried out a qualitative analysis in order to detect the main performance problems. We find problems concerning segmentation rules and concerning FreeLing. The main problem of segmentation rules concerns situations where the element que (“that”) is involved at the same time that the conjunction y (“and”). Example 3a shows DiSeg segmentation and example 3b shows the correct segmentation.
3a. [El perfil del usuario sería el de un varón (51,4%) de mediana edad (43,2 años) que consulta por patología traumática (50,5%) y procede de la comarca sanitaria cercana al hospital.]  

ENGLISH TRANSLATION OF 3a. [The general profile of users would be a man (51.4%) of middle age (43.2 years) who consults because of traumaticopathologies (50.5%) and comes from the sanitary area near the hospital.]  

One of the DiSeg segmentation rules indicates that the relative que ("that") is not considered as segmentation boundary. Nevertheless, another of these rules indicates that, if there is a coordinative conjunction (like y ("and")) and next there is a verb, that conjunction constitutes a possible segmentation boundary. Thus, DiSeg does not segment before que ("that"), but it segments just before y ("and"), because it finds the verb procede ("comes from") before the end of the sentence. We have detected several cases with a similar problem.  

Moreover, we detect two errors due to a wrong sentence segmentation of FreeLing. Example 4 shows one of them (example 4a shows DiSeg segmentation and example 4b shows the correct segmentation).  

4a. [No encontramos cambios en la medición del ángulo astrágalo-calcáneo en AP. Realizamos una descripción de nuestra serie y una discusión acerca de la técnica y de la indicación actual de la cirugía en esta patología.]  

ENGLISH TRANSLATION OF 4a. [In the measurement of the talar-calcaneal angle in AP there were no changes. We carry out a description of our series and a discussion about the surgical technique and the present indication in this pathology.]  

4b. [No encontramos cambios en la medición del ángulo astrágalo-calcáneo en AP.]  

ENGLISH TRANSLATION OF 4b. [In the measurement of the talar-calcaneal angle in AP there were no changes.]  

The sentence segmentation module does not segment correctly these two sentences, probably because it considers “AP” as an abbreviation and it does not detect the beginning of the second sentence. This problem causes an error in the discourse segmentation of DiSeg.  

In order to solve the main detected systematic segmentation errors, we developed three more symbolic rules. These three rules imply the use of que ("that") and y ("and"). Using DiSeg, the segmentation of example 5a is obtained:  

5a. [La gastroplastia vertical anillada, técnica de inicio puramente restrictiva, fue abandonada debido a resultados inadecuados en cuanto a la pérdida de peso y vómitos repetidos que condicionaban la calidad de vida.]  

ENGLISH TRANSLATION OF 5a. [Vertical banded gastroplasty, a purely restrictive procedure, was abandoned due to inadequate results in weight loss and repetitive vomiting that inferred with oral intake and social life.]  

This segmentation is wrong. The system detects the conjunction y ("and") and the verb condicionaban ("inferred") and it interprets that this passage has to be segmented into two EDUs. Nevertheless, using DiSeg 1.0, (including the new three rules), the same passage is segmented correctly, as example 5b shows:  

5b. [La gastroplastia vertical anillada, técnica de inicio puramente restrictiva, fue abandonada debido a resultados inadecuados en cuanto a la pérdida de peso y vómitos repetidos que condicionaban la calidad de vida.]  

As we have mention in Section 4, the new version of the system (DiSeg-1.0), including these new rules, improves the results.  

With regard to the experiment over the terminological sub-corpus, we have detected that the DiSeg_1.0 errors are basically due to two problems. On the one hand, the system has difficulties to detect the elements of the Same-Unit relation, that is, an EDU that is “broken” and contains another EDU inside. Example 6a shows the segmentation obtained by DiSeg_1.0 and example 6b shows the correct segmentation.  

6a. [En décadas precedentes se ha puesto de manifiesto, y así lo han atestiguado muchos investigadores de la terminología científica serbia, una tendencia a importar préstamos de unidades estructurales tanto léxicas como otras mayores del inglés a una serie de registros científicos específicos.]  

ENGLISH TRANSLATION OF 6a. [In the recent past, a trend has been noted, and it has been attested by many researchers in the area of Serbian scientific terminology, of importing borrowings of lexical and larger structural units from English into specific scientific registers.]  

6b. [En décadas precedentes se ha puesto de manifiesto, y así lo han atestiguado muchos investigadores de la terminología científica serbia, una tendencia a importar préstamos de unidades estructurales tanto léxicas como otras mayores del inglés a una serie de registros científicos específicos.]  

On the other hand, there were problems with regard to the sense disambiguation of FreeLing, which caused some segmentation errors. For example, in Spanish, the word doctrina ("doctrine") could be a noun or a verb. In the example 7, this word constitutes a noun, but FreeLing assigns it the category of verb. Because of this, DiSeg 1.0 detects the conjunction y ("and") and the supposed verb doctrina ("doctrine"), so it segments an EDU wrongly, as example 7a shows. Example 7b includes the correct segmentation.  

7a. [Las mismas presentan fundamentalmente tres tipos de documentos que responden a las tres fuentes principales del Derecho: jurisprudencia, legislación y doctrina de los autores.]  

ENGLISH TRANSLATION OF 7a. [They contain basically three types of documents, corresponding to the three main sources of law: jurisprudence, legislation and the doctrine of authors.]  

7b. [Las mismas presentan fundamentalmente tres tipos de documentos que responden a las tres fuentes principales del Derecho: jurisprudencia, legislación y doctrina de los autores.]
6. Conclusions

We have developed DiSeg, the first discourse segmenter for Spanish, based on lexical and syntactic rules. We consider that this research constitutes an important step into the research on automatic discourse parsing in Spanish, because there are not many works about this topic for this language. We have evaluated DiSeg performance, measuring precision, recall and F-Score, comparing it with a gold standard that we have carried out. Performance is good if we compare it with the baseline segmenters. Moreover, results are similar to the ones obtained by Tofiloski et al. (2009). Additionally, we think that the gold standard we have carried out is a good contribution in order to encourage other researchers to go on investigating in this field.

As future work, we plan to solve the detected errors, using more symbolic rules and/or machine learning approaches like in the work of Afantenos et al. (2010). We would like to carry out some experiments using the tags of the Freeing dependency parser instead the tags of the shallow parser and lexical units. We think some of the rules we have designed would be more effective using this type of syntactic analysis. Moreover, we will apply DiSeg to another Spanish corpus including general texts from the Wikipedia. The final goal of the project is to develop the first discourse parser for Spanish on an open platform, easily adaptable to the other Latin languages implemented in FreeLing. We would remark that the methodology we have used in this work is a global approach that would be employed to develop discourse segmenters for other Latin languages. As we have explained, this segmentation is carried out using shallow parsing (with Freeing, an open-source multilingual platform) and some specific lexical tags (mostly discourse markers), so this strategy could be adapted to other similar languages.

Acknowledgments

We recognize the distinction and support provided by Sociedad Mexicana de Inteligencia Artificial (SMIA) and the 9th Mexican International Conference on Artificial Intelligence (MICAI-2010) in order to enhance, improve, and publish this work. This research is partially supported by: a postdoctoral Grant (National Program for Mobility of Research Human Resources; National Plan of Scientific Research, Development and Innovation 2008-2011) given to Iria da Cunha by Ministerio de Ciencia e Innovación, Spain; the research project CONACyT (Mexico), number 82050; the research project PAPIIT-DCAPA, number IN403108, and the Spanish Science and Innovation Ministry through the project KNOW2 (TIN2009-14715-C04-03), RICOTERM (FFI2010-21365-C03-01) and APLE (FFI2009-12188-C05-01). Thanks Daniel Owen for the English corrections of this paper.

Appendix A

In this section we show an example of DiSeg_1.0 segmentation.

A.1. Original Spanish passage

Con el fin de predecir la tasa esperable de ganglios centinela en nuestra población, hemos analizado la tasa de invasión axilar en los últimos 400 casos de cáncer de mama pT1 operados por nosotros, utilizando la técnica clásica de linfadenectomía axilar completa. De los 400 tumores 336 (84.0%) fueron carcinomas ductales infiltrantes NOS, 32 (8.0%) carcinomas lobulillares, 22 carcinomas tubulares puros (5.5%), y los 10 restantes correspondieron a otras variedades histológicas menos frecuentes. A la hora de realizar el estudio del ganglio centinela en cánceres de mama T1 en nuestra población, cabe esperar globalmente la detección de un ganglio positivo en al menos una de cada cuatro pacientes.

A.2. English translation given by the text author

In order to predict the expectable positive sentinel node rate in our population, we analyzed the rate of axillary invasion in the last 400 pT1 breast cancer cases operated by us, using the classical technique of complete axillary dissection. Of the 400 tumors, 336 (84.0%) were ductal NOS infiltrating carcinomas, 32 (8.0%) lobular carcinomas, the remaining 10 belonging to other, less frequent histological varieties. When studying the sentinel node in T1 breast cancers in our population, the detection of a positive node may globally be expected in one out of four patients.

DiSeg_1.0 segmentation.

In this section we show a screenshot of the complete DiSeg_1.0 xml output.
References


