Discourse Segmentation for Spanish Based on Shallow Parsing

Iria da Cunha[1,2,3], Eric SanJuan[2], Juan-Manuel Torres-Moreno[2,4], Marina Lloberes[5], and Irene Castellón[5]

[1] Institute for Applied Linguistics (UPF), C/Roc Boronat 138, Barcelona, Spain
[2] Laboratoire Informatique d'Avignon, BP1228, 84911, Avignon Cedex 9, France
[3] Instituto de Ingeniería (UNAM), Ciudad Universitaria, 04510, Mexico
[4] École Polytechnique de Montréal/DGI, Montréal (Québec), Canada
[5] GRIAL – Universitat de Barcelona, C/Gran Via de les Corts 585, Barcelona, Spain
iria.dacunha@upf.edu,
{eric.sanjuan,juan-manuel.torres}@univ-avignon.fr,
{marina.lloberes,icastellon}@ub.edu

Abstract. 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, obtaining promising results.

Keywords: Discourse Parsing, Discourse Segmentation, Rhetorical Structure Theory.

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 [1,2], Japanese [3] and Brazilian Portuguese [4,5]. Most of them use the framework of Rhetorical Structure Theory (RST) [6]. 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 [7]: "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 [7[1],8], Brazilian Portuguese [9][2] and French [22], but not for Spanish. All of them require some syntactic analysis of the sentences. [7,8,9] rely on a set of linguistic rules. [22] relies on machine learning techniques: it learns rules automatically from thoroughly annotated texts.

[1] http://www.sfu.ca/~mtaboada/research/SLSeg.html
[2] http://www.nilc.icmc.usp.br/~erick/segmenter/

G. Sidorov et al. (Eds.): MICAI 2010, Part I, LNAI 6437, pp. 13–23, 2010.

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 that require 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 in [7,22], we evaluate 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 implementation. The gold standard corpus and the experimental setup is detailed in section 4, meanwhile results are reported in section 5. Finally, section 6 is devoted to conclusions and future work.

2 Methodology

The theoretical framework of our research is based on Rhetorical Structure Theory (RST), as defined in [6]. As mentioned in [10], this theory is used in a wide range of NLP applications like automatic text generation [11,12,13], summarization [1,14,15], translation [16,17], etc. In all these applications, RST is used to obtain a deeper linguistic analysis. Figure 1 shows an example of RST discourse tree (with three relations: Background, Contrast and Concession).

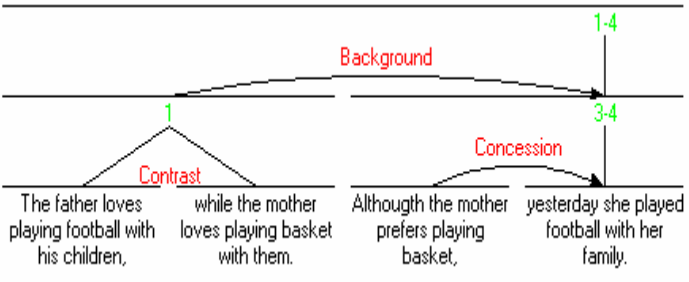

Fig. 1. Example of a RST discourse tree

As it can be seen in this example, RST can work at two levels. At sentence level to analyse 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 [23] 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 this 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 [18], but with some differences, similar to those included in [7] and [19]. 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 [20]);
- 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 Implementation

DiSeg implementation relies on the open source software FreeLing [21] for the Part of Speech (PoS) and shallow parsing. This open-source highly scalable resource for NLP

applications is based on simple Hidden Markov Model (HMM) classifiers and on readable optimal Context Free Grammars (CFG), which can be easily adapted to specific needs. Therefore, we carry out some modifications into the default grammar of the shallow parser. These were mainly recategorizations of some elements (as prepositions, prepositional phrases, adverbs or adverbial phrases) into discourse markers (*disc_mk)*. FreeLing output is then encoded into an XML structure to be processed by perl programs that apply 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 (recategorizated elements) and grup_sp_inf (infinitives). The only text markers that are used apart from these tags are the period and two words: "that" (*que*) and "for" (*para*).

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. Figure 2 shows DiSeg architecture.

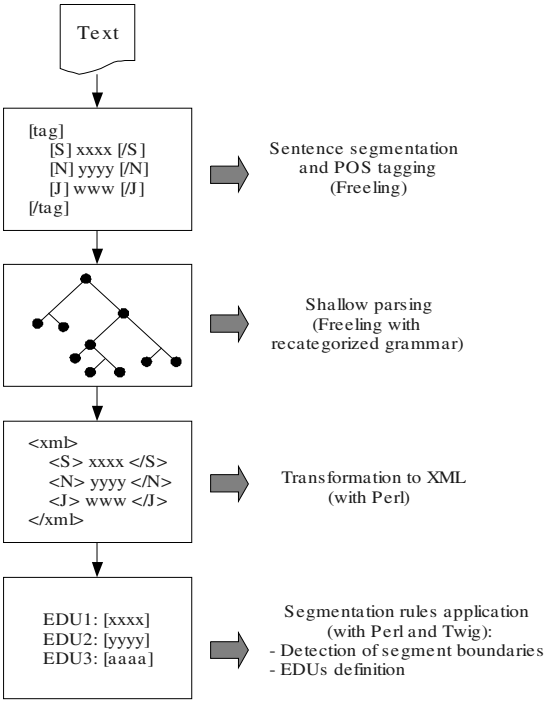

Fig. 2. DiSeg architecture

DiSeg can be used on-line at http://diseg.termwatch.es. The system is also available under General Public License (GPL). 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 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*).

The XML elements "<seg rule="rule1">...<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 extend the non CFG module is necessary.

4 Experiments and Evaluation

The gold standard test corpus 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)[3]. The corpus includes 169 sentences. Text average is 8.45 sentences. The longest text contains 21 sentences and the shortest text contains 3 sentences. The corpus contains 3981 words. Text average is 199.05 words. The longest text contains 474 words and the shortest text contains 59 words. This corpus includes 203 EDUs. Text average is 10.15 EDUs (the maximum 28 EDUs and the minimum 3 EDUs). These statistics are similar to the statistics of the gold standard corpus used in [7] to develop a discourse segmenter for English, obtaining very good results.

This 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 annotators, a consensus was reached. We use the consensual segmentation as gold standard. This gold standard is also available at http://diseg.termwatch.es.

We ran DiSeg over this corpus for evaluation and computed precision, recall and F-Score measures among detected and correct boundaries. Precision is the number of correct boundaries detected by the system over the total number of detected ones. Recall is the same number of correct boundaries detected by the system but divided

[3] http://www.gacetamedicabilbao.org/web/es/

this time by the total number of real boundaries existing in the gold standard corpus. As in [7], we do not count sentence boundaries in order to not inflate the results.

For this evaluation, we used three baseline segmenters:

1. *Baseline_0* only considers sentences as EDUs. This is not a trivial baseline since its precision is 100% by definition and four texts in the gold standard have no other type of EDUs.
2. *Baseline_1* inserts discourse boundaries before each *coor* tag introduced by the Freeling shallow parsing.
3. *Baseline_2* considers both tags indicating *coor* and *conj_subord* but only the last segment at the right of the sentence with a verb is considered as an EDU.

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.

5 Results

Table 1 contains the results of the evaluation. Results show that *DiSeg* full system outperforms *DiSeg-base* and all the baselines. 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, appears to give the best results).

These results are similar to those obtained by the discourse segmenter for English introduced in [7]: 93% of precision, 74% of recall and 83% of F-Score. Thus, we consider that DiSeg results are promising.

Table 1. Results of the evaluation

	Precision	**Recall**	**F-Score**
DiSeg	**71%**	**98%**	**80%**
DiSeg-base	70%	88%	74%
Baseline_2	68%	82%	72%
Baseline_1	33%	70%	39%
Baseline_0	100%	49%	62%

After a quantitative analysis of the results, we carry 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%)]*$_{EDU1}$ [*y procede de la comarca sanitaria cercana al hospital.*]$_{EDU2}$

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 traumatologic pathologies (50.5%)]$_{EDU1}$ [and comes from the sanitary area near the hospital.]$_{EDU2}$

3b. [*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.*]$_{EDU1}$

ENGLISH TRANSLATION OF 3B. [The general profile of users would be a man (51.4%) of middle age (43.2 years) who consults because of traumatologic pathologies (50.5%) and comes from the sanitary area near the hospital.]$_{EDU1}$

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.*]$_{EDU1}$

ENGLISH TRANSLATION OF 4A. [In the measurement of the talar-calcaneal angle in AP there was no changes. We carry out a description of our series and a discussion about the surgical technique and the present indication in this pathology.]$_{EDU1}$

4a. [*No encontramos cambios en la medición del ángulo astrágalo-calcáneo en AP.*]$_{EDU1}$ [*Realizamos una descripción de nuestra serie y una discusión acerca de la técnica y de la indicación actual de la cirugía en esta patología.*]$_{EDU2}$

ENGLISH TRANSLATION OF 4B. [In the measurement of the talar-calcaneal angle in AP there was no changes.]$_{EDU1}$ [We carry out a description of our series and a discussion about the surgical technique and the present indication in this pathology.]$_{EDU2}$

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.

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 in [7]. 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 [22]. 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.

Acknowledgments. This work 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, number 82050, the research project PAPIIT-DGAPA, number IN403108, and the research project "Representación del Conocimiento Semántico" (SKR) KNOW2 (TIN2009-14715-C0403).

References

1. Marcu, D.: The Theory and Practice of Discourse Parsing Summarization. Institute of Technology, Massachusetts (2000a)
2. Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. Computational Linguistics 26(3), 395–448 (2000b)
3. Sumita, K., Ono, K., Chino, T., Ukita, T., Amano, S.: A discourse structure analyzer for Japonese text. In: International Conference on Fifth Generation Computer Systems, pp. 1133–1140 (1992)
4. Pardo, T.A.S., Nunes, M.G.V., Rino, L.M.F.: DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. In: Bazzan, A.L.C., Labidi, S. (eds.) SBIA 2004. LNCS (LNAI), vol. 3171, pp. 224–234. Springer, Heidelberg (2004)
5. Pardo, T.A.S., Nunes, M.G.V.: On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. Journal of Theoretical and Applied Computing 15(2), 43–64 (2008)
6. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), 243–281 (1988)
7. Tofiloski, M., Brooke, J., Taboada, M.: A Syntactic and Lexical-Based Discourse Segmenter. In: 47th Annual Meeting of the Association for Computational Linguistics, Singapur (2009)
8. Soricut, R., Marcu, D.: Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In: 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada, pp. 149–156 (2003)
9. Mazeiro, E., Pardo, T.A.S., Nunes, M.G.V.: Identificação automática de segmentos discursivos: o uso do parser PALAVRAS. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC). São Carlos, São Paulo (2007)

10. Taboada, M., Mann, W.C.: Applications of rhetorical structure theory. Discourse Studies 8(4), 567–588 (2005)
11. Hovy, E.: Automated discourse generation using discourse structure relations. Artificial Intelligence 63, 341–385 (1993)
12. Dale, R., Hovy, E., Rösner, D., Stock, O.: Aspects of Automated Natural Language Generation. Springer, Berlin (1992)
13. O'Donnell, M., Mellish, C., Oberlander, J., Knott, A.: ILEX: An architecture for a dynamic Hypertext generation system. Natural Language Engineering 7, 225–250 (2001)
14. Radev, D.: A common theory of information fusion from multiple text sources. Step one: Cross document structure. In: Dybkjær, L., Hasida, K., Traum, D. (eds.) 1st SIGdial Workshop on Discourse and Dialogue, Hong-Kong, pp. 74–83 (2000)
15. Pardo, T.A.S., Rino, L.H.M.: DMSumm: Review and assessment. In: Ranchhod, E., Mamede, N.J. (eds.) PorTAL 2002. LNCS (LNAI), vol. 2389, pp. 263–274. Springer, Heidelberg (2002)
16. Ghorbel, H., Ballim, A., Coray, G.: ROSETTA: Rhetorical and Semantic Environment for Text Alignment. In: Rayson, P., Wilson, A., McEnery, A.M., Hardie, A., Khoja, S. (eds.) Proceedings of Corpus Linguistics, Lancaster, pp. 224–233 (2001)
17. Marcu, D., Carlson, L., Watanabe, M.: The automatic translation of discourse structures. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000), Seattle, vol. 1, pp. 9–17 (2000)
18. Carlson, L., Marcu, D.: Discourse Tagging Reference Manual. ISI Technical Report ISITR-545. University of Southern California, Los Angeles (2001)
19. da Cunha, I., Iruskieta, M.: La influencia del anotador y las técnicas de traducción en el desarrollo de árboles retóricos. Un estudio en español y euskera. In: 7th Brazilian Symposium in Information and Human Language Technology (STIL). Universidade de São Paulo, São Carlos (2009)
20. Alonso, L.: Representing discourse for automatic text summarization via shallow NLP techniques. PhD thesis. Universitat de Barcelona, Barcelona (2005)
21. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L.l., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: 5th International Conference on Language Resources and Evaluation. ELRA (2006)
22. Afantenos, S., Denis, P., Muller, P., Danlos, L.: Learning Recursive Segments for Discourse Parsing. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010)
23. da Cunha, I., Fernández, S., Velázquez-Morales, P., Vivaldi, J., SanJuan, E., Torres-Moreno, J.-M.: A New Hybrid Summarizer Based on Vector Space Model, Statistical Physics and Linguistics. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 872–882. Springer, Heidelberg (2007)

APPENDIX A. Example of DiSeg segmentation

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.

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.

Segmented text

<rst>
****
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.
****
</rst>

APPENDIX B. Screenshot of the complete DiSeg XML output

```
<XML>
    <S>
      <seg rule="rule1">
        <disc_mk>
          <term cat="SPS00" lem="con">Con</term>
          <term cat="DA0MS0" lem="el">el</term>
          <term cat="NCMS000" lem="fin">fin</term>
          <term cat="SPS00" lem="de">de</term>
        </disc_mk>
        <grup_verb_inf>
          <infinitiu role="head">
            <inf role="head">
              <forma_inf role="head">
                <term cat="VMN0000" lem="predecir">predecir</term>
              </forma_inf>
            </inf>
          </infinitiu>
        </grup_verb_inf>
          <grup_nom_fs role="head">
            <n_fs role="head">
              <term cat="NCFS000" lem="población">población</term>
            </n_fs>
[...]
          </grup_nom_fs>
        </sn>
        <segD rule="rule4.2">
          <term cat="Fc" lem=",">,</term>
        </segD>
        <subseg>
          <grup_verb>
            <verb role="head">
              <vaux>
                <term cat="VAIP1P0" lem="haber">hemos</term>
              </vaux>
              <parti role="head">
                <term cat="VMP00SM" lem="analizar">analizado</term>
              </parti>
            </verb>
          </grup_verb>
          <sn>
            <espec_fs>
              <j_fs role="head">
                <term cat="DA0FS0" lem="el">la</term>
              </j_fs>
            </espec_fs>
            <grup_nom_fs role="head">
              <n_fs role="head">
                <term cat="NCFS000" lem="tasa">tasa</term>
              </n_fs>
            </grup_nom_fs>
          </sn>
[...]
```