Comparación de dos métodos para la extracción de opiniones en textos en español

Aiala Rosa, Dina Wonsever, Jean-Luc Minel

▶ **To cite this version:**

Aiala Rosa, Dina Wonsever, Jean-Luc Minel. Comparación de dos métodos para la extracción de opiniones en textos en español. IBERAMIA 2010, Nov 2010, Bahia Blanca, Argentina. pp.99 ,108. halshs-00785391

HAL Id: halshs-00785391
https://shs.hal.science/halshs-00785391v1

Submitted on 6 Feb 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparación de dos métodos para la extracción de opiniones en textos en español

Aiala Rosá[1,2], Dina Wonsever[1], Jean-luc Minel[2],

[1] Facultad de Ingeniería, Universidad de la República
J. Herrera y Reissig 565, Montevideo, Uruguay
[2] Modyco, Université Paris Ouest Nanterre
200, avenue de la République, Nanterre, France
{aialar, wonsever}@fing.edu.uy
jean-luc.minel@u-paris10.fr

Resumen. En este artículo abordamos el problema de la identificación de opiniones en textos en español y nos concentramos en la comparación de dos tipos de métodos para la extracción de la fuente: reglas desarrolladas manualmente y aprendizaje, sobre un corpus anotado, de clasificadores secuenciales. El primer sistema alcanza un 87% de medida-F. Para el segundo sistema desarrollado se obtiene un 81% de medida-F. Es necesario hacer pruebas más extensivas pero una primera reflexión indica que con el método de aprendizaje se obtienen rápidamente resultados de un nivel comparable, aunque aprovechando en este caso el análisis del dominio y los recursos léxicos recopilados que fueron necesarios para elaborar las reglas.

Palabras clave: Extracción de información, Análisis de opiniones, *Conditional Random Fields*

1 Introducción

Actualmente es muy común la lectura de prensa electrónica a través de Internet, por lo que resulta de utilidad contar con sistemas para la extracción de diferente tipo de información según las necesidades de los usuarios. En los textos de prensa, es usual encontrar citas a palabras de otros participantes, por lo que existen dos aspectos que son particularmente importantes: poder asociar un punto de vista expresado en el texto con la fuente apropiada y asignar un valor afectivo (positivo, negativo, neutro) a la opinión citada. Existen incluso aplicaciones comerciales que brindan este tipo de servicio (http://www.jodange.com), realizando análisis de prensa en inglés.

En este artículo abordamos el problema de la identificación de opiniones en textos en español y nos concentramos en la comparación de dos tipos de métodos: reglas desarrolladas manualmente y aprendizaje, sobre un corpus anotado, de clasificadores secuenciales. No incluimos en este trabajo el estudio del contenido afectivo de la opinión.

Entendemos por opinión la reproducción de un acto verbal en el cual un enunciador se pronuncia sobre algún tema (ej. *El investigador de la Politécnica*

afirma que el principal problema de este sistema es conseguir que sea fácil de usar), o cualquier mención a creencias o posturas de participantes del discurso (ej. *El Pri acepta participar en el debate*) distintos del autor del texto. Caracterizamos una opinión mediante cuatro elementos, no necesariamente explícitos siempre en toda opinión: el predicado, normalmente se trata de verbos como *decir*, *opinar*, *aceptar*, *rechazar* (en los ejemplos: *afirma* y *acepta*); la fuente, participante del discurso al cual se puede atribuir la opinión (en los ejemplos: *El investigador de la Politécnica* y *El Pri*); el asunto, tema sobre el cual se opina (en el segundo ejemplo: *participar en el debate*) y el mensaje, contenido de la opinión (en el primer ejemplo: *que el principal problema de este sistema es conseguir que sea fácil de usar*).

Existen numerosos trabajos que abordan estas temáticas y que constituyen antecedentes de nuestra propuesta. En [13] se analizan en detalle diversos conceptos del área "Opinion Mining" o "Sentiment Analysis" y se presentan las propuestas, los recursos y las aplicaciones más importantes.

Para nuestra propuesta, que se centra en la identificación de los predicados de opinión, la fuente, el asunto y el mensaje, hemos considerado fundamentalmente: el esquema para anotación de emociones y opiniones de [17]; el trabajo sobre identificación de la fuente y del contenido proposicional de la opinión (similar a lo que hemos denominado mensaje) propuesto en [3]; el sistema para identificación de la fuente utilizando métodos estadísticos de [5]; el método de extracción de la fuente y el tópico (lo que nosotros llamamos asunto) de [9]; el estudio sobre identificación de la fuente y el asunto presentado en [14], el trabajo de anotación de tópico de [15] y la identificación de fuente y tópico propuesta en [11].

Si bien existen algunos trabajos para el español en esta área, en general abordan el problema de la orientación semántica en textos, centrándose en muchos casos en la construcción de diccionarios afectivos [2, 4]. No tenemos conocimiento de la existencia de sistemas de identificación de los componentes de la opinión, tal como aquí los definimos, para el español.

En lo que sigue, presentamos brevemente el modelo que definimos para representar la opinión y dos métodos informáticos para su reconocimiento automático.

En primer lugar, describimos un sistema para la identificación de las opiniones y sus componentes, basado en reglas deducidas a partir de la inspección de un corpus y la incorporación de recursos léxicos. El sistema alcanza un 85% de medida-F para la identificación de opiniones completas. Más adelante presentamos también los valores obtenidos para cada componente de la opinión.

El segundo sistema consiste en la aplicación de técnicas de aprendizaje automático para la identificación de las fuentes de opiniones. Se evaluó sobre un corpus de testeo obteniéndose un 81% de medida-F. Sobre este mismo corpus de testeo se volvió a evaluar la identificación de fuentes del sistema de reglas, esta vez filtrando posibles errores provenientes de la etapa de identificación de predicados, obteniéndose un 87% de medida-F.

Los valores anteriores fueron tomados considerando tanto el reconocimiento exacto como el reconocimiento parcial de los elementos. Más adelante presentamos también los valores de reconocimiento exacto exclusivamente.

2 Elementos que componen la opinión

Modelamos la opinión en base a la presencia de un predicado de opinión y de sus argumentos característicos. Dentro del conjunto de predicados de opinión, se encuentran, en primer lugar, los verbos pertenecientes a diversas clases semánticas: verbos de comunicación (*decir, declarar*), creencia (*creer, opinar*), valoración (*criticar, felicitar*), aceptación (*aceptar, rechazar*) y sensación (*gustar, molestar*). Dentro del conjunto de predicados de opinión, también incluimos nombres, en general derivados de los verbos anteriores, como *opinión, declaración, apoyo*. Por último, consideramos predicados de opinión las preposiciones o locuciones prepositivas como *según, de acuerdo con*, etc.

Los argumentos que identificamos como relevantes para los predicados de opinión son la fuente, el asunto y el mensaje. Para establecer este esquema nos basamos en los esquemas sintáctico-semánticos propuestos en ADESSE[1] para los verbos de las clases antes mencionadas [7] y en algunos de los frames de Spanish FrameNet[2] [16].

2.1 Algunos ejemplos

En un caso típico de discurso reproducido como (1), observamos un verbo de comunicación que constituye el predicado de la opinión. El sujeto del verbo es la fuente y la proposición subordinada que contiene lo expresado por la fuente es el mensaje. Normalmente, no hay un segmento que corresponda a lo que llamamos asunto.

(1) [El investigador de la Politécnica]$_f$ [afirma]$_p$ [que el principal problema de este sistema es conseguir que sea fácil de usar]$_m$.

Por otro lado, en (2), en vez del mensaje tenemos en la oración una mención al asunto. Se trata de un verbo que introduce un discurso referido en el cual, a diferencia de lo que sucede en el discurso reproducido, sólo se hace mención a la existencia del acto de enunciación, sin que haya una reproducción de las palabras emitidas, o sea, del mensaje [12].

(2) [El abogado de Fernando Botero]f [habló]p [sobre el tema]a con Semana.

De todos modos, encontramos también casos de discurso reproducido en los cuales se incluye una mención al asunto de la opinión, como en (3), y casos de discurso referido en los cuales se citan las palabras emitidas, como en (4).

(3) [Sobre la partitura]$_a$ [Ros Marbá]$_f$ [afirma]$_p$ [que es "enormemente teatral. Se define a los personajes desde la propia música, cada uno tiene ....]$_m$

(4) En una carta escrita por Dalí en Neuilly en abril de 1951, [el artista]$_f$ [habla]$_p$ [sobre su divina inspiración]$_a$: ["Yo quería que el próximo Cristo que pintase .... "]$_m$.

Como dijimos anteriormente, la opinión puede estar expresada por una nominalización o un inciso con según. En estos casos, el predicado en vez de ser un verbo es un nombre (5) o una preposición (6).

1 http://webs.uvigo.es/adesse/

2 http://gemini.uab.es:9080/SFNsite

(5) [La Iglesia Católica]$_f$ también hizo pública su [opinión]$_p$, [contraria a cualquier tipo de despenalización]$_a$.

(6) [Este sistema se utiliza en Estados Unidos desde 1982]$_m$, [según]$_p$ [Roque Pifarré]$_f$.

3 Sistema de reglas

Desarrollamos un sistema compuesto por reglas que permite la identificación de los diferentes elementos de la opinión. El sistema toma como entrada un texto pre-procesado por un POS-tagger, Freeling [1] y posteriormente por el sistema Clatex [19] que segmenta las oraciones en proposiciones. Luego se aplican varios módulos de reglas para el reconocimiento de opiniones, obteniéndose como salida el texto con anotaciones en formato xml que muestran las opiniones encontradas y los elementos que las componen. La salida del sistema se ilustra en el siguiente ejemplo:

<opinion> <mensaje>"Botnia en Uruguay está teniendo un comportamiento excelente" </mensaje>, <predicado>dijo</predicado> <fuente>el ministro de Medio Ambiente, Carlos Colacce</fuente> </opinion>

Las reglas que integran el sistema se basan en el formalismo de reglas contextuales definido por [18], incluyendo algunas extensiones posteriores. Este tipo de regla permite la especificación de contextos, zonas de exclusión, operadores de opcionalidad, negación, eliminación de etiquetas ya existentes, entre otros. Además, cada regla ofrece la posibilidad de chequear condiciones diversas de los elementos que la componen, como por ejemplo, la pertenencia a una lista de palabras, funcionalidad imprescindible para este trabajo.

Las reglas diseñadas para este sistema se agrupan en módulos según el elemento que reconocen: predicado (27 reglas), fuente (42 reglas), asunto (22 reglas), mensaje (8 reglas). Hay, además, un módulo final que arma la opinión completa (37 reglas) y algunos módulos de reglas accesorias (37 reglas), como el módulo que arma grupos nominales complejos del tipo [*El director del Hospital Maciel, Daniel Parada*] o los módulos que identifican y combinan elementos subjetivos y los operadores que actúan sobre ellos.

3.1 Recursos léxicos

Algunas de las reglas, principalmente las que marcan los predicados, se basan fuertemente en recursos léxicos: lista de verbos y nombres de opinión, lista de indicadores de persona (*señor, doctor*), lista de indicadores de institución (*institución, hospital*), lista de verbos soporte (*realizar, emitir*) y lista de preposiciones o locuciones prepositivas que introducen el asunto (*sobre, en cuanto a, respecto a*).

En particular, las listas de verbos y nombres de opinión fueron creadas manualmente en base al análisis del corpus de desarrollo, que contiene textos de prensa en español provenientes de: Corin [8], Corpus del español [6] y un corpus de prensa digital confeccionado para este trabajo. Con el fin de minimizar las

ambigüedades, sólo se incorporaron a las listas los verbos y nombres con usos frecuentes en contextos de opinión, para lo cual se consultaron las listas de las clases correspondientes de los recursos léxicos ADESSE y Spanish FrameNet. Al momento de la evaluación que presentamos a continuación, contábamos con 128 predicados de opinión (86 verbos y 42 nombres). La información asociada a cada elemento incluye propiedades sintácticas y semánticas.

3.2 Evaluación del sistema de reglas

Para la evaluación del sistema trabajamos con un corpus de prensa digital, tomado de las mismas publicaciones de las cuales provienen los textos del corpus de desarrollo. El corpus de evaluación tiene aproximadamente 13.000 palabras y contiene un total de 302 opiniones.

Se aplicaron las reglas al corpus completo y se hizo una revisión manual de la salida con el objetivo de evaluar la identificación de cada elemento de la opinión y, además, la identificación de opiniones completas.

Además de realizarse la evaluación de la performance de las reglas, durante la etapa de revisión manual se corrigió el texto anotado de modo que se cuenta actualmente con un corpus de 13.000 palabras anotado correctamente con las opiniones y sus componentes.

La tabla 1 muestra los valores de *Precision*, *Recall* y *Medida F* calculados.

Tabla 1: Resultados de la evaluación del sistema de reglas. Las filas representan: PR: *Precision*, REC-E: *Recall* considerando solo recuperación exacta, REC-P: *Recall* considerando también recuperación parcial, F: Medida-F calculada considerando el valor de REC-P.

	pred	**fuen**	**asun**	**mens**	**opinión**
PR	92 %	93 %	96 %	95 %	94 %
REC-E	91 %	63 %	45 %	58 %	42 %
REC-P	91 %	72 %	62 %	84 %	77 %
F	91.5 %	81 %	75 %	89 %	85 %

Los resultados muestran que la cobertura de predicados de opinión es bastante elevada (91% de *recall*). En una etapa posterior a esta evaluación, las listas de predicados se extendieron en base a diferentes revisiones de corpus, llegando a 170 predicados. Actualmente estamos trabajando en la incorporación de más predicados, tomados de los recursos léxicos mencionados. Queda pendiente una nueva evaluación del sistema, con un nuevo corpus, para evaluar en qué medida mejoran los resultados al crecer las listas de predicados.

Se observa una cantidad importante de mensajes reconocidos en forma parcial, en algunos casos debido a errores en la segmentación de las proposiciones por parte de Clatex.

Es importante señalar que la cantidad de opiniones con menciones explícitas al asunto (introducido por expresiones como *sobre*, *en cuanto a* , etc.) es muy baja, 74

en un total de 302 (24%), por lo que la evaluación del reconocimiento de este elemento no es muy significativa.

4 Sistema de aprendizaje automático

El segundo sistema implementado aplica aprendizaje automático en base a *Conditional Random Fields* (CRF) [10], modelo probabilístico secuencial. En un problema de clasificación (nuestro caso), en el modelo se estima la probabilidad condicional de una secuencia de valores de la variable de salida, dada una secuencia de entrada. Es posible usar un amplio espectro de funciones tipo atributo construidas a partir de los datos de entrada, sin requerimientos de independencia para distintos atributos entre sí.

Hasta el momento, el sistema solo se desarrolló para el reconocimiento de fuentes de opiniones.

4.1 Descripción de los experimentos

Aplicamos CRF al problema de identificación de las fuentes de las opiniones mediante la herramienta CRF++[3].

Utilizamos con un corpus de entrenamiento de 30.000 palabras y un corpus de testeo de 10.000. Usamos 8 atributos (palabra, lema, categoría gramatical, tipo de nombre, tipo de verbo, número, género, indicador de introducción de fuente), además de una columna final con las anotaciones de salida (B-FU, comienzo de fuente; I-FU, interior a fuente; O, otro). Algunos de los atributos tienen el significado usual, los definidos especialmente para este problema se explican a continuación:

- 'tipo de nombre' permite indicar, para las palabras cuya categoría es nombre, si se trata de un predicado de opinión (como *declaración*, *apoyo*, etc.) o de cualquier otro nombre, en cuyo caso se indica si es nombre común o nombre propio. Para las categorías restantes este atributo vale 0.
- 'tipo de verbo' permite indicar, para las palabras cuya categoría es verbo, si se trata de un predicado de opinión (como *declarar*, *apoyar*, etc.) o de cualquier otro verbo, en cuyo caso se indica si es un verbo principal, un verbo auxiliar o un verbo soporte (*emitir*, *realizar*, etc.). Para las categorías restantes este atributo vale 0.
- 'introductor de fuente' indica si el elemento es un introductor de fuente preposicional como *según*, *de acuerdo con*, etc. En caso afirmativo el atributo vale 1, en otros casos vale 0.

Cabe aclarar que, si bien los predicados de opinión verbales y nominales fueron tomados de las listas utilizadas por el sistema de reglas, se hizo además una revisión manual de los corpus para detectar elementos relevantes no pertenecientes a las listas. Esto se hizo con el objetivo de poder evaluar los resultados de aplicar CRF al reconocimiento de fuentes, filtrando en lo posible errores provenientes de factores externos a esta tarea concreta, como la no identificación de algunos predicados de opinión.

3 http://crfpp.sourceforge.net/

En la figura 1 mostramos algunos fragmentos del corpus de entrenamiento. Hay tres predicados de opinión: *destacó*, *precisiones* y *según*. El primero, por ser verbal, lleva el valor 'Op' en la columna 5, correspondiente al atributo 'tipo de verbo'. El segundo, por ser nominal, lleva el valor 'Op' en la columna 4, correspondiente al atributo 'tipo de nombre'. El tercero es un introductor de fuente, por lo que lleva el valor '1' en la columna 8. Las fuentes a identificar son las secuencias de palabras [*La ministra*], [*La ministra de Salud Pública*] y [*Muñoz*]. A estas secuencias de palabras de les asignan los valores de salida (columna 9) B-FU, para el primer elemento de cada secuencia, e I-FU, para los restantes. Todas las demás líneas del ejemplo tienen O en la última columna, indicando que la palabra no pertenece a ninguna fuente.

La operativa de CRF++ requiere que se definan familias de funciones de atributos (*templates*, según la terminología de la documentación).

Para nuestro trabajo especificamos diferentes *templates* según el atributo:

- para los atributos 'categoría', 'tipo de nombre' y 'tipo de verbo', para cada línea del corpus se combinan los valores tomados por el atributo en esa línea con los valores tomados en las cuatro líneas anteriores y en las cuatro siguientes.
- para los atributos palabra, lema, número, género e introductor de fuente, para cada línea del corpus se combinan los valores tomados por el atributo en esa línea con los valores tomados en las dos líneas anteriores y en las dos siguientes.

Para algunos atributos se definieron ventanas más grandes, hasta 4 líneas anteriores y siguientes, por ser los atributos que más inciden en la determinación de la fuente.

```
...
La el D 0 0 S F 0 B-FU
ministra ministra N C 0 S F 0 I-FU
destacó destacar V 0 Op S 0 0 O
la el D 0 0 S F 0 O
prevención prevención N C 0 S F 0 O
...
La el D 0 0 S F 0 B-FU
ministra ministra N C 0 S F 0 I-FU
de de S 0 0 0 0 0 I-FU
Salud_Pública salud_pública N P 0 0 0 0 I-FU
realizó realizar V 0 Sop S 0 0 O
algunas alguno D 0 0 P F 0 O
precisiones precisión N Op 0 P F 0 O
...
Las el D 0 0 P F 0 O
medidas medida N C 0 P F 0 O
, , Fc 0 0 0 0 0 O
según según S 0 0 0 0 1 O
Muñoz muñoz N P 0 0 0 0 B-FU
, , Fc 0 0 0 0 0 O
tuvieron tener V 0 M P 0 0 O
...
```

Fig. 1. Extracto del corpus de entrenamiento

4.2 Evaluación

El corpus de entrenamiento tiene un total de 486 fuentes y el de testeo tiene 158. Luego de obtenerse un modelo a partir del aprendizaje aplicado sobre el corpus de entrenamiento, se aplica dicho modelo al corpus de testeo. Se obtiene como salida el corpus con anotaciones, las cuales deben compararse con las anotaciones manuales para así poder evaluar los resultados.

De dicha comparación surgen los valores de *Precision* y *Recall* que se muestran en la tabla siguiente. Al igual que como se hizo con el sistema de reglas, se calcularon dos valores para cada medida, por un lado elementos correctos exactos y por otro lado elementos correctos parciales.

Tabla 2: Resultados de la evaluación del sistema basado en CRF.

	Precision	*Recall*	*F*
exacto	92%	64%	75.5%
parcial	99%	69%	81%

5 Comparación de resultados de los dos métodos para la identificación de fuente

Para poder comparar los dos sistemas realizamos una nueva evaluación del sistema de reglas, esta vez solamente para el reconocimiento de fuentes, asumiendo que todos los predicados de opinión son reconocidos, como lo hicimos con el sistema de aprendizaje automático. Para esto agregamos a las listas que utilizan las reglas algunos verbos y nombres que no pertenecían a ellas (obteniéndose la lista extendida que mencionamos en 3.2).

La tabla 3 muestra los resultados obtenidos para la identificación de fuentes por los dos sistemas.

Tabla 3: Comparación de los resultados de los dos sistemas.

	Precision		*Recall*		*F*	
	exacto	parcial	exacto	parcial	exacto	parcial
reglas	89%	99%	70%	78%	78%	87%
CRF	92%	99%	64%	69%	75.5%	81%

Describiremos algunos casos que consideramos interesantes para comentar. La fuente introducida por *según* en el texto *según informaron a Montevideo Portal fuentes locales* es reconocida correctamente por el sistema de reglas pero no por el sistema basado en CRF. Creemos que una posible causa para la falla del segundo

sistema es la existencia de un segmento entre el predicado de opinión (*según*) y la fuente (*fuentes locales*), situación que se repite en otros fragmentos.

Entre los casos de fuentes reconocidas solo por el sistema basado en CRF, destacamos fuentes que contienen algunas palabras no contempladas por las reglas de formación de grupos nominales candidatos a ser fuentes, por ejemplo, la conjunción *y* en [*El ministro de Vivienda y Medio Ambiente*, *Carlos Colacce*], *dijo....* o la preposición *por* en [*La fuente consultada porObserva*] *añadió que ...*

6 Conclusiones

Hemos analizado la expresión de las opiniones en textos en español e implementado el reconocimiento de las mismas. En el caso de las fuentes de las opiniones, hemos experimentado con dos tipos de métodos: elaboración manual de reglas y aprendizaje modelando por *Conditional Random Fields*. Los resultados de ambos métodos son, en primera instancia, similares. Si bien el sistema de reglas insumió un tiempo de desarrollo mucho mayor, permitió conocer las peculiaridades lingüísticas del dominio de las opiniones y acumular recursos que consideramos útiles para el sistema basado en CRF. Estimamos de todos modos que es necesario ampliar los experimentos antes de mayores conclusiones.

Los valores de medida-F obtenidos son muy alentadores si tomamos como referencia algunos trabajos cercanos al nuestro, aunque para otros idiomas y con algunas diferencias en cuanto al tipo de fuente que se busca (78.4% para el chino [11] y 69.4% para el inglés [5], valores calculados incluyendo elementos reconocidos en forma parcial).

Deseamos señalar como un aporte de nuestro trabajo al desarrollo de herramientas para el procesamiento de textos en español, la creación de varios recursos lingüísticos que pueden ser reutilizados en otros contextos, además de los sistemas para el reconocimiento de opiniones y sus fuentes.

Referencias

1. Atserias, J., B. Casas, E. Comelles, M. González, L. Padró y M. Padró. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. En Proceedings of the fifth international conference on Language Resources and Evaluation (LREC) ELRA. (2006).
2. Banea, Carmen, Rada Mihalcea, Janyce Wiebe, Samer Hassan. Multilingual Subjectivity Analysis Using Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNLP). (2008).
3. Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou y Dan Jurafsky. Automatic extraction of opinion propositions and their holders. En AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004).
4. Brooke, J., M. Tofiloski y M. Taboada. Cross-Linguistic Sentiment Analysis: From English to Spanish. RANLP 2009, Recent Advances in Natural Language Processing. Borovets, Bulgaria. (2009).

5. Choi, Yejin, Claire Cardie, Ellen Riloff y Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. En Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (Vancouver, British Columbia, Canada). Human Language Technology Conference. Association for Computational Linguistics. (2005).
6. Davies, Mark. Corpus del español (100 millones de palabras, siglo XIII - siglo XX). Disponible actualmente en http://www.corpusdelespanol.org. (2002).
7. García-Miguel, J., L. Costas y S. Martínez. Diátesis verbales y esquemas construccionales. Verbos, clases semánticas y esquemas sintácticosemánticos en el proyecto ADESSE. Entre semántica léxica, teoría del léxico y sintaxis, 373-384. (2005).
8. Grassi, Mariela, Marisa Malcuori, Javier Couto, Juan José Prada y Dina Wonsever. Corpus informatizado: textos del español del Uruguay (CORIN), SLPLT-2 - Second International Workshop on Spanish Language Processing and Language Technologies - Jaén, España. (2001).
9. Kim, Soo-Min y Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. En Proceedings of the Workshop on Sentiment and Subjectivity in Text (Sydney, Australia, July 22 - 22, 2006). ACL Workshops. Association for Computational Linguistics, Morristown, NJ, 1-8. (2006).
10. Lafferty, J., A. McCallum y F. Pereira. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* En Proc. of ICML, pp.282-289. (2001).
11. Lu, Bin. Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts. En Proceedings of the NAACL HLT 2010 Student Research Workshop, pages 46–51, Los Angeles, California, June 2010. (2010).
12. Maldonado, Concepción. Discurso directo y discurso indirecto. En Ignacio Bosque y Violeta Demonte, Gramática descriptiva de la lengua española (Entre la oración y el discurso. Morfología), 3549- 3596. (1999).
13. Pang, Bo y Lillian Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135. (2008).
14. Ruppenhofer, Josef, Swapna Somasundaran y Janyce Wiebe. Finding the Sources and Targets of Subjective Expressions. The Sixth International Conference on Language Resources and Evaluation (LREC 2008). (2008).
15. Stoyanov, Veselin y Claire Cardie. Annotating Topics of Opinions. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco. (2008).
16. Subirats-Rüggeberg, Carlos y Miriam R. L. Petruck. Surprise: Spanish FrameNet! En E. Hajicova, A. Kotesovcova & Jiri Mirovsky (eds.), Proceedings of CIL 17. CD-ROM. Prague: Matfyzpress. (2003).
17. Wiebe, Janyce, Theresa Wilson y Claire Cardie. Annotating expressions of opinions and emotions in language. En Language Resources and Evaluation (formerly Computers and the Humanities), 39(2- 3):165210. (2005).
18. Wonsever, Dina y Jean-Luc Minel. Contextual Rules for Text Analysis. En Lecture Notes in Computer Science. (2004).
19. Wonsever, Dina, Serrana Caviglia, Javier Couto y Aiala Rosá. Un sistema para la segmentación en proposiciones de textos en español. In Letras de hoje 144 (41). (2006).