Marcadores discursivos del español. Descripción y propuesta de detección automática

Discourse markers in the Spanish language. Description and proposal of automatic detection

Walter Koza
Grupo INFOSUR-UNR-Becario de CONICET
Rosario, Argentina
Walter_koza@yahoo.com.ar

Abstract

This article presents a method for the automatic detection of discourse markers in the Spanish language. Discourse markers are marginal elements which establish relationships among text segments with the aim of signalling and organizing the reading. The recognition of these constructions is of great importance for several computational lingüistics tasks. Nevertheless, one of the problems possibly developing is that of ambiguous constructions, that is to say, syntagms acting as discourse markers in some given contexts but not in others.
Punctuation leads an important role to solve this kind of drawbacks and Prada puts forward a series of disambiguation hypotheses, putting emphasis upon it.
The implantation-in-machine work is based on such hypotheses.
On this occasion, software Smorph was employed. Smorph is a text analyser and generator which performs the previous delimitation of the segments to be considered and the morphological analysis, all in only one stage.
With the detection method proposed, a 98% precision and a 97% coverage were achieved.
Key words: Discourse markers, Computational Lingüístics, Punctuation, Comma, Smorph.

Resumen

En este artículo se presenta un método de detección automática de los marcadores discursivos del español. Los marcadores discursivos son elementos marginales que establecen relaciones entre segmentos textuales con el objetivo de guiar y ordenar la lectura. El reconocimiento de estas construcciones es de gran importancia para varias tareas de la lingüística computacional. No obstante, uno de los problemas que puede suscitarse es el caso de construcciones ambiguas, es decir, sintagmas que actúan como marcadores discursivos en determinados contextos y en otros no.
La puntuación cumple un papel importante para solucionar este tipo de inconvenientes y Prada formula una serie de hipótesis de desambiguación en las que hace hincapié en ella. El trabajo de implantación en máquina está basado en dichas hipótesis.
En esta ocasión, se recurrió al software Smorph. Smorph es un analizador y generador textual que en una única etapa realiza la delimitación previa de los segmentos a considerar y el análisis morfológico.
Con el método de detección propuesto, se logró una precisión del 98% y una cobertura del 97%.

Palabras claves: Marcadores discursivos, Lingüística Computacional, Puntuación, Coma, Smorph.

0. INTRODUCCIÓN

En este artículo se presenta la descripción y una propuesta de reconocimiento automático de los marcadores discursivos, mediante la utilización de la herramienta informática Smorph [1]. Para ello, se van a tomar los lineamientos de la tesis doctoral que en estos momentos estoy realizando sobre el análisis de la puntuación en el marco de la lingüística computacional, dirigida por la Doctora Zulema Solana y financiada por una beca de CONICET.

La lingüística computacional pertenece tanto al terreno de la lingüística aplicada como así también al de la Inteligencia Artificial y tiene por objetivo el análisis y la producción del lenguaje natural. Esta disciplina comenzó a desarrollarse en la segunda mitad del siglo pasado y estaba orientada hacia la obtención de traductores automáticos. Posteriormente, y con el avance y la popularización de la informática, que permite, pero a la vez exige, el acceso a grandes masas textuales con los fines más variados, se impuso la necesidad de compatibilizar los lenguajes naturales y los lenguajes computacionales con el propósito de posibilitar el tratamiento de estos grandes corpus lingüísticos. A tales efectos, en estos días, la lingüística computacional no es solo un medio alternativo de procesamiento del lenguaje, sino que, por el contrario, se ha convertido en un recurso imprescindible. [2]

Con respecto a los estudios sobre este campo realizados en el país, en la Facultad de Humanidades y Artes de la UNR, durante los años 2004 y 2005, la Maestría de Teoría Lingüística y Adquisición del Lenguaje ha organizado cursos de lingüística computacional dictados por el Doctor Grabriel Bès, de la Universidad Blaise Pascal (Clermont-Ferrant, Francia), y la Doctora Zulema Solana, con el asesoramiento científico de este, ha elaborado el proyecto “INFOSUR. Investigación y desarrollo” (PID de la CECYT UNR), que en estos días trabaja en diversos proyectos vinculados. Tanto los cursos realizados, como la participación en el proyecto INFOSUR, han permitido el acceso al programa Smorph. Este software es un analizador y generador textual que, en una sola etapa, presenta la segmentación de los términos a reconocer (una palabra o una expresión del tipo ‘sin embargo’, ‘es decir’, etcétera) y el análisis morfológico mediante la asignación de una etiqueta morfosintáctica para cada uno de esos segmentos (‘nombre’, ‘verbo’, ‘punto’, ‘coma’, etcétera).

Smorph es una herramienta declarativa, lo que implica que la información utilizada está separada de la maquinaria algorítmica. Esto hace que se la pueda adaptar al uso que quiera darse, ya que con el mismo software se puede tratar cualquier lengua si le cambia la información lingüística.

De acuerdo con la descripción de su autor, Aït-Mokhtar, este programa es un utilitario que agrupa un conjunto de funcionalidades en torno a la morfología: compilación de diccionario, análisis y generación morfológicas, segmentación y lematización de textos. Además, permite construir diccionarios electrónicos voluminosos, necesarios para un análisis lingüístico de textos y que a su vez pueden utilizarse en generación o en segmentación y análisis. [1]

Vale aclarar que Smorph es un programa de prueba que no está a la venta ni tiene licencia de uso, se pudo acceder a él gracias a la gentileza del Doctor Gabriel Bès.

Con respecto a la puntuación, es pertinente señalar la escasez de estudios y la poca atención que ha recibido; ya sea desde el punto de vista del análisis del discurso, de la relación con la enseñanza aprendizaje, o bien, del aspecto semántico y pragmático.
No obstante, este sistema se presenta como elemento clave al momento de analizar un escrito. La puntuación no deviene en un sistema rígido e invariable como el de la ortografía, sino que por el contrario, sus normas son menos rígidas y, en gran parte, están sujetas a la idea estilística del escritor. Prada [3] la concibe como un sistema de signos gráficos cuyo propósito es delimitar unidades de procesamiento del texto, lo que minimizaría el esfuerzo del lector en la comprensión. Esto significa que los signos de puntuación constituyen un mecanismo para organizar un texto; delimita las llamadas unidades textuales y, por lo tanto, los cambios o no de tema.
Sobre la base de este planteo y a partir de los trabajos de Nunberg [4] y Figueras [5], se plantean cinco categorías textuales básicas:

- Párrafo: dado por punto y aparte;
- Enunciado textual: dado por punto y seguido;
- Cláusula textual: dada por punto y coma;
- Enunciado oracional: dado por dos puntos;
- Sintagma: dado por la coma.

La segmentación textual es una fase necesaria para una gran variedad de tareas realizadas en el tratamiento automático; ya sea para el análisis sintáctico, el resumen automático, el filtrado de textos, etcétera. No obstante, varios autores señalan que esta tarea no está adecuadamente tratada. Las herramientas que existen en el mercado como segmentadores-balizadores de textos como HTML utilizan, para textos bien estructurados, balizas hasta el punto y aparte; pero la segmentación de los textos en unidades menores ("frases") representa una tarea que actualmente no está bien definida. [6]
En mi trabajo actual, me propongo seguir avanzando dentro de estas cuestiones en la medida en que creo que el reconocimiento adecuado de las marcas de puntuación en relación con la segmentación textual, y a partir de las funciones gramaticales que presentan, constituiría un aporte al desarrollo del análisis automático.
Mi análisis se centra en la coma, por ser el signo de puntuación más complejo y el que mayor número de funciones posee. Una tarea inicial fue poder establecer una clasificación de sus funciones lingüísticas. Para ello, se cotejaron diferentes clasificaciones de la coma, realizadas por diversos autores; entre ellos, puede mencionarse a Alcoba [7], Simone [8] y Figueras [5]. A partir de estos antecedentes, se estableció una clasificación propia basada en criterios puramente sintácticos y gramaticales, pertinente para la posterior modelización y creación de reglas.
Las funciones lingüísticas de la coma son las siguientes:

I- Separar los términos en las enumeraciones.

(1) [Compró pan, verduras y frutas.]

II- Indicar elipsis.

(2) [Juan lee un libro; María, una revista.]

III- Delimitar vocativos.

(3) [Me gustaría, Juan, que vayamos al cine.]

IV- Delimitar construcciones dicendi, introductoras del discurso citado.

(4) [“Mi esposa”, dijo Juan, “ha tenido amoríos con otro hombre”.]

V- Delimitar construcciones interjectivas.

(5) [No sé, ¡ay de mí!, si lo soportaré.]

VI- Señalar alteraciones del orden regular.

(6) [En el parque, un señor vestido de payaso regala caramelos.]

VII- Delimitar construcciones que modifican a un sintagma de la cláusula, como por ejemplo aposiciones (7), subordinadas relativas explicativas (8), etcétera.

(7) [María, la esposa de Juan, ha sido una mujer infiel.]
(8) [Juan, que no era tonto, comprendió al instante lo que sucedía.]

VIII- Especificar marcadores discursivos y expresiones conjuntivas.

(9) [No lograron acordar, o sea, se separaron.]

En relación con el último ítem, la coma sirve, en algunos casos, para determinar si una construcción es o no un marcador discursivo. Por ejemplo, puede observarse las diferentes funciones que cumplen las expresiones “por un lado” y “por otro lado” en (10) y (11).

> (10) [**Por un lado,** se prevé una "redistribución de los suministros de gas" que tienen contratados las centrales térmicas. (…)
> Y, **por otro lado**, a partir de 2007 se procederá a la desregulación total de los suministros y cualquier usuario residencial podrá contratar la provisión de energía con el generador o comercializador que más le convenga.] (*Clarín*, 29/04/04)
> (11) [La verdad va **por un lado** y la política **por otro lado**.]

En el primero de los ejemplos, se tratan de marcadores discursivos estructuradores de la información, cuya función es asociar la información de los dos párrafos, dada por la ilación ‘por un lado’, ‘por otro lado’. En cambio, en (11) las mismas construcciones actúan como complementos verbales.

En esta ocasión, se propone un método de detección automática de aquellos marcadores discursivos que requieren de coma para evitar ambigüedades. El artículo se organiza de la siguiente manera:

En primer lugar, se presenta una descripción de los marcardores discursivos y la clasificación que de ellos presentan Zorraquino y Portolés Lázaro [9]. Luego se traerán a colación los antecedentes sobre el tratamiento de estas construcciones en la lingüística computacional, haciendo hincapié en el trabajo de Prada [3]. En tercer lugar, se describe la implantación en máquina y los resultados obtenidos. Por último, se presentan las conclusiones derivadas de la investigación.

1. ACERCA DE LOS MARCADORES DISCURSIVOS

1.1. Consideraciones generales

Los marcadores discursivos aluden a un conjunto de términos que establecen relaciones entre segmentos textuales con el objetivo de guiar y ordenar los procesos de interpretación en la comprensión de textos. A comienzos de la década del setenta, dos disciplinas nacientes, la lingüística textual y la pragmática, se centraron en el estudio de estos términos en la medida en que confirmaban sus respectivos puntos de partida: la propuesta de romper las fronteras de la oración como límite último de los estudios del lenguaje. [10]
Estos elementos textuales han recibido varias denominaciones, “enlaces extraoracionales” [11], “ordenadores del discurso” [12], “conectores pragmáticos” [13], “conectores extraoracionales” [14], “operadores discursivos” [15], “ordenadores del discurso” [16], “conectores enunciativos” [17], “conectores pragmáticos” [18], “marcadores del discurso” [9], “marcadores y conectores” [19], etcétera.
Para tratar estos términos, me baso en los estudios de Zorraquino y Portolés, quienes se refieren a ellos con la denominación de “marcadores del discurso” y los definen de la siguiente manera:

> “Los ‘marcadores del discurso’ son unidades lingüísticas invariables, no ejercen una función sintáctica en el marco de la predicación oracional –son, pues, elementos marginales– y poseen un cometido coincidente en el discurso: el de guiar, de acuerdo con sus distintas propiedades morfosintácticas, semánticas y pragmáticas, las inferencias que se realizan en la comunicación”. [9]

Una cuestión a considerar es que los marcadores discursivos presentan una variada gama de particularidades y poseen una complejidad que escapa a todo intento exhaustivo de sistematización. No obstante, estos autores tratan de presentar una clasificación de ellos, basada en dos condiciones:
a) los elementos agrupados deben compartir propiedades gramaticales homogéneas (los marcadores tratados se ajustan, en general, a las categorías tradicionales de los adverbios, las locuciones adverbiales y ciertas interjecciones);
b) a su vez, sus características semánticas (la forma de significar o configurar su significado) deben ser las propias de los marcadores discursivos; esto es, que no presentan un contenido referencial o denotador sino que muestran un significado de procesamiento.
Estas construcciones son de gran ayuda en el proceso de inferencias durante la comprensión lectora. Las inferencias son un conjunto de operaciones de razonamiento que realiza el lector en el momento de leer. Zorraquino y portolés solo van a considerar como marcadores a aquellas expresiones que no contribuyan al significado conceptual de los enunciados, sino que orienten y ordenen las inferencias que cabe obtener de ellos.

Esto quiere decir que los marcadores contribuyen al procesamiento de lo que se comunica y no a la representación de la realidad comunicada.

1.2. Clasificación de los marcadores discursivos

Zorraquino y Portolés distinguen cinco grupos de marcadores; el primero se denomina "estructuradores de la información" y se los utiliza para señalar la organización informativa de los discursos. Estos carecen de significado argumentativo y se dividen en:

- Comentadores: introducen un nuevo comentario ("pues", "pues bien", "así las cosas");
- Ordenadores: agrupan varios miembros del discurso como partes de un único comentario. Estos marcadores, por lo general, están basados en la numeración ("primero", "segundo"), en lo espacial ("por un lado" "por otro lado", "por una parte" "por otra parte") o en lo temporal ("después", "luego", "finalmente"). Algunos de ellos forman pares correlativos que pueden estar seguidos por un tercer miembro también con ordenador: "por un lado/por otro (lado)", "por una parte/por otra (parte)", etcétera. Los ordenadores, a su vez se clasifican en tres tipos:

  1) Marcadores de apertura: abren una serie en el discurso ("en primer lugar", "primeramente", "por un lado");
  2) Marcadores de continuidad: indican que el miembro que acompañan forma parte de una serie de la cual no son el elemento inicial ("en segundo/tercer/…/lugar", "por otra parte", "por otro lado");
  3) Marcadores de cierre: señalan el fin de una serie discursiva ("por último", "en último lugar", "finalmente");

- Disgresores: introducen un comentario lateral respecto de la planificación del discurso anterior ("por cierto", "a todo esto", "a propósito").

El segundo grupo de marcadores discursivos es el de los "conectores", que vinculan semántica y pragmáticamente a un miembro del discurso con otro anterior, de tal forma que el marcador guía las inferencias que se efectúan del conjunto de los dos miembros discursivos conectados. Pueden reconocerse tres grupos:

- Conectores aditivos: unen a un miembro anterior con otro de su misma orientación ("además", "encima", "aparte", "incluso");
- Conectores consecutivos: conectan a un consecuente con su antecedente ("por tanto", "por consiguiente", "por ende", "en consecuencia");
- Conectores contraargumentativos: eliminan algunas de las conclusiones que pudieran inferirse de un miembro anterior ("en cambio", "por el contrario", "sin embargo", "no obstante").

El tercer grupo de marcadores del discurso es el de los “reformuladores”. Estos presentan a un miembro del discurso como una expresión más adecuada de lo que se pretendió decir con un miembro precedente. Los reformuladores se subdividen en cuatro grupos, a saber:

- Reformuladores explicativos: presentan un nuevo miembro del discurso como una explicación anterior (“o sea”, “esto es”, “es decir”);
- Reformuladores rectificativos: corrigen a un miembro discursivo anterior (“mejor dicho”, “mejor aún”, “más bien”);
- Reformuladores de distanciamiento: privan de pertinencia al miembro discursivo anterior (“en cualquier caso”, “en todo caso”, “de todos modos”);
- Reformuladores recapitulativos: introducen una recapitulación o conclusión de un miembro discursivo anterior o una serie de ellos (“en suma”, “en conclusión”, “en definitiva”, “en fin”, “al fin y al cabo”).

El cuarto grupo es el de los “operadores argumentativos”. En este caso, el marcador condiciona, por su significado, las posibilidades argumentativas del miembro en que se incluye, sin relacionarlo con otro anterior. Pueden establecerse dos grupos de operadores argumentativos:

- Operadores de refuerzo argumentativo: su significado refuerza como argumento el miembro del discurso en el que se encuentra frente a otros posibles argumentos (“en realidad”, “en el fondo”, “de hecho”);
- Operadores de concreción: muestran el miembro del discurso en el que se localizan como una concreción o un ejemplo de una generalización (“por ejemplo”, “en particular”).

Finalmente el quinto grupo es el que remite a los marcadores “conversacionales”. Respecto de ellos, los autores aclaran que con esta división no se pretende determinar un límite estricto entre lo conversacional y lo no conversacional, puesto que:

> “todo discurso es, en esencia, dialógico y, de hecho, mucho de los marcadores que se han incluido en los grupos precedentes pueden aparecer también en la conversación; asimismo, bastantes marcadores conversacionales se emplean a menudo en los textos escritos”. [9]

Sin embargo, la conversación constituye una situación comunicativa particular y con propiedades específicas que van a determinar o favorecer la presencia de ciertos marcadores. Los marcadores conversacionales se distribuyen a en cuatro grupos:

- De modalidad epistémico: señalan el grado de certeza, de evidencia, etcétera, que el hablante atribuye al miembro o a los miembros del discurso con el que se vincula cada partícula (“claro”, “por lo visto”, “desde luego”);
- De modalidad deóntica: indican diversas actitudes volitivas del hablante respecto del miembro o miembros del discurso en el que aquellos comparecen (“bueno”, “bien vale”);

- Enfocadores de la alteridad: orientan sobre la forma como el hablante se sitúa en relación con su interlocutor en la interacción comunicativa (“hombre”, “mira”, “oye”);
- Metadiscursivos conversacionales: sirven para estructurar la conversación; es decir para distinguir bloques informativos, por ejemplo, o para alternar o mantener los turnos de palabra, etcétera (“bueno”, “eh”, “este”).

En el presente trabajo no se tratará a este grupo de marcadores en la medida en que no hace a los objetivos que es el análisis de textos escritos.

2. SOBRE LOS MARCADORES DISCURSIVOS EN EL ÁMBITO DE LA LINGÜÍSTICA COMPUTACIONAL

Los marcadores discursivos son de gran utilidad en varias tareas del procesamiento del lenguaje natural, como el auto resumen, la traducción automática, análisis sintáctico, etcétera, ya que aportan información muy rica sobre la estructura discursiva, con un bajo coste de procesamiento. No obstante, señalan Alonso, Castellón y Padró [20], la comunidad científica no ha alcanzado el consenso respecto de su delimitación y caracterización.

> “Esta falta de consenso se debe, por un lado, a la preeminencia de las aproximaciones de tipo deductivo, con un sesgo importante por una teoría subyacente y, por otro, a la subordinación de la mayor parte de caracterizaciones a una tarea computacional concreta, lo que suele conllevar soluciones *ad hoc*”. [20]

Con respecto a los antecedentes, entre los trabajos sobre estas construcciones, pueden mencionarse los realizados por Dale y Knott [21], quienes proponen mecanismos formales para la detección y sistematización de estas unidades. Knott [22] aplica estos mecanismos para obtener y caracterizar un conjunto de unos 200 marcadores discursivos del inglés. Sobre la base de este trabajo, Marcu [23] desarrolla un sistema de análisis de la estructura retórica para el inglés basado en la información discursiva que obtiene de un conjunto de 400 marcadores discursivos.
Para Alonso, Castellón y Padró, las falencias de los estudios mencionados radican en que no solucionan la creación de un listado extenso y no controvertido de los marcadores discursivos para uso computacional o cómo abordar la creación de estos recursos para otras lenguas.
Para el caso del español, estos autores presentan una construcción de un lexicón computacional de marcadores discursivos, implementado en un sistema de resumen genérico de extracción, que puede funcionar autónomamente o en colaboración con otras técnicas o tipos de información. El sistema se compone de dos módulos que utilizan el lexicón:

- El **segmentador**, que detecta las unidades discursivas básicas y
- El **interpretador**, que pondera cada segmento según su relevancia discursiva.

Dicho lexicón fue implementado en un sistema de ayuda para el resumen automático. La primera etapa del trabajo fue realizada mediante métodos empíricos e implantación

en un sistema de resumen. Posteriormente, presentan mejoras estructurales y de contenido mediante la aplicación de técnicas clustering.
Otra de las investigaciones realizadas sobre los marcadores discursivos es la realizada por Prada [3], quien tomando como base la clasificación propuesta por Zorraquino y Portolés, desarrolla una propuesta de implantación en máquina de una serie de reglas que permiten el reconocimiento automático de los marcadores. Los signos de puntuación cumplen un papel importante para solucionar problemas de ambigüedad como los descriptos en (10) y (11) y Prada formula una serie de hipótesis de desambiguación en las que hace hincapié en ellos. Tales hipótesis son las siguientes:

Hipótesis 1: Ordenadores
Se podrá considerar estructura ordenadora a aquella en la que cada marcador de la secuencia sea el primer elemento en una oración o en un párrafo y además debe estar seguida de una coma. En el caso de ocurrencias intraoracionales (fundamentalmente para los de continuidad) la coma o una conjunción son obligatorias antes del marcador. Por otro lado, para que un marcador ordenador cumpla la función discursiva de continuidad en una secuencia, debe estar acompañado de algún otro marcador del conjunto.

Hipótesis 2: Adverbios y marcadores de continuidad
Para que un marcador discursivo de continuidad que a la vez es adverbio (‘igualmente’, ‘asimismo’) se comporte como un ordenador debe aparecer:

- Entre comas si es intraoracional;
- Si ocupa el primer lugar en una oración debe venir seguido de coma (a excepción de “igualmente”);
- El resto de los ordenadores de continuidad de este grupo, solamente aparecen al comienzo de una oración y deben ir seguidos de coma.

Hipótesis 3: Marcador entre signos de puntuación
Para que el marcador cumpla la función discursiva debe aparecer:

- Entre comas si es intraoracional;
- Si ocupa el primer lugar en una oración debe ir seguido de coma.

Hipótesis 4: Marcador que no ocupa el último lugar de una frase o proposición
Para que cumpla la función discursiva, nunca puede ser el último término de una oración o proposición.

Hipótesis 5: Marcador que exige signo de puntuación previo
Para que cumpla la función discursiva, debe aparecer:

- Una coma o punto y coma previo a él si es intraoracional;
- Si ocupa el primer lugar en una oración, la puntuación puede faltar.

Hipótesis 6: Marcador que exige signo de puntuación posterior
Para que cumpla la función discursiva, debe aparecer:

- Una coma luego del término si es intraoracional;
- Si ocupa el primer lugar en una oración, la puntuación puede faltar.

Basándome en estas hipótesis, elaboré una lista de marcadores discursivos que cargué en las entradas de Smorph con el objetivo de que el programa efectuase el reconocimiento automático sin que presente las ambigüedades que podrían suscitarse con estos términos.

3. IMPLANTACIÓN EN MÁQUINA

Se trabajó con un texto de diez mil palabras conformado por artículos periodísticos del corpus del equipo Infosur. La herramienta trabaja con un archivo de entradas. Las entradas constituyen el diccionario lingüístico en el que palabras y signos de puntuación tienen la posibilidad de aparecer, ya sea a partir de los lemas con la indicación precisa del modelo morfológico al que pertenecen [24], o bien, como se ha hecho con los marcadores discursivos, directamente con la indicación de los rasgos morfológicos.
La ventaja de este programa es que se pueden declarar todo tipo de expresión, ya sea una sola palabra como un conjunto de estas. A sí pues, se tienen entradas de nombres, adjetivos, verbos, preposiciones, adverbios, etcétera. Asimismo, se creó una nueva categoría de entradas correspondiente a los marcadores discursivos. Estos se cargaron incluyendo las comas, los puntos y comas y copulativos que pudieran necesitarse a fin de eliminar las ambigüedades.
Con el propósito de ilustrar, se presentan algunos de ellos tal como fueron cargados en las entradas de Smorph:

```
en_primer_lugar,        /marcd/estruc/orden/ap  .
y_en_primer_lugar,      /marcd/estruc/orden/ap  .
,_en_primer_lugar,      /marcd/estruc/orden/ap  .
en_segundo_lugar,       /marcd/estruc/orden/cont  .
por_añadidura,      /marcd/conect/aditiv  .
por_lo_tanto,       /marcd/conect/consec  .
en_efecto,      /marcd/refor/exp  .
a_saber:        /marcd/refor/exp  . [25]
```

Una vez que elaborada la lista con los marcadores discursivos, se puede proceder al análisis morfológico de cadenas textuales. He aquí ejemplos del reconocimiento realizado:

De todos modos, afirmó que todavía no puede anticipar (…)
'De todos modos,'.
['de_todos_modos,', 'EMS', 'marcd', 'TMARCD', 'refor', 'CLREF', 'distan'].
'afirmó'.
['afirmar', 'EMS', 'svn'].
(…)

Sin embargo, los funcionarios que lo visitaron ayer (…)
'Sin embargo,'.
['sin_embargo,', 'EMS', 'marcd', 'TMARCD', 'conect', 'CLCON', 'contra'].
'los funcionarios'.
['el funcionario', 'EMS', 'SN'].

(…)

(…) *estos programas, en efecto, es que contradicen la ética* (…)
', en efecto,'.
[',_en_efecto,', 'EMS', 'marcd', 'TMARCD', 'refor', 'CLREF', 'exp'].
'es'.
['ser', 'EMS', 'svn'].
(…)

No importa cuánto valga, en definitiva, el esfuerzo de una persona.
', en definitiva,'.
[',_en_definitiva,', 'EMS', 'marcd', 'TMARCD', 'refor', 'CLREF', 'recap'].
'el esfuerzo de una persona'.
['el esfuerzo de una persona', 'EMS', 'SN'].

Por medio de este método de detección, se logró un resultado de un 98% de precisión y un 97% de cobertura.

4. CONSIDERACIONES FINALES

Se presentó un breve panorama de la lingüística computacional, haciendo hincapié en la importancia de la puntuación en el análisis automático. Por otro lado, se propuso una clasificación de las funciones de la coma basada en criterios gramaticales y sintácticos. Asimismo, se planteó el papel fundamental que cumple este signo de puntuación en el momento de evitar construcciones ambiguas.
Luego se presentó una descripción de los marcadores discursivos y la clasificación que de ellos realizaron Zorraquino y Portolés y se trajeron a colación los antecedentes del tratamiento de estas construcciones en la lingüística computacional.
A partir de las hipótesis de Prada para considerar o no a una construcción, marcador discursivo, se cargó en el archivo de entradas de Smorph, una serie de marcadores clasificados según los criterios de Zorraquino y Portolés.
Se probó la implantación realizada en un corpus de diez mil palabras y los resultados obtenidos fueron de un 98% de precisión y un 97% de cobertura.

Referencias

[1] Aít-Mokhtar, S. L’analyse présintaxique en une seule étape. Tesis doctoral. Universidad Blaise-Pascal/GRIL, Clermont-Ferrand, 1998.
[2] Para una reseña de la Lingüística Computacional, cf. Bonino, R. “Presentación de la Lingüística Computacional”, en Solana Zulema (Ed.) La Interlengua de los aprendientes del español como L2. Aportes de la Lingüística Informática. Centro de Estudios de Adquisición del Lenguaje, Facultad de Humanidades y Artes, UNR, Rosario 2009.
[3] Prada, J. Marcadores del discurso en español. Análisis y representación. Tesis de Maestría, Universidad de la República, Montevideo, 2001.
[4] Nunberg, G. The Linguistics of Punctuation. En CSLI Lecture Notes. Stanford University Press. Stanford, 1990.
[5] Figueras, C. “La semántica procedimental de la puntuación”. En Espéculo. Revista de estudios literarios. Universidad Complutense de Madrid. Madrid, 1997.
[6] Mourad, G. “La segmentation de textes par l’etude de la ponctuation”. En Acte de Colloque International, CIDE ’99. Document Electronique Dynamique, pp. 155-171. Damas, Syrie, 1999.

[7] Alcoba, S. “Puntuación y melodía de la frase”. En Alcoba (coord.) La expresión oral. Ariel Practicum. Madrid, 2000.
[8] Simone, R. “Riflessioni sulla virgola”. En Orsolini, M. y Pontecorvo, C. La costruzione del testo nei bambini. Firence, 1991.
[9] Martín Zorraquino, M. y Portolés Lázaro, J. “Los marcadores del discurso”. En Bosque, I. y Demonte V. (Dirs.), Gramática descriptiva de la lengua española, Tomo III. Espasa Calpe. Madrid, 1999.
[10] Portolés Lázaro, J. Marcadores del discurso. Ariel. Barcelona, 1998.
[11] Gili Gaya, S. Curso superior de sintaxis española. Bibliograf. Barcelona, 1943.
[12] Alcina, F. y Blecua, J. Gramática española. Ariel. Barcelona, 1975.
[13] Stubbs, M. Análisis del discurso. Alianza. Madrid, 1983.
[14] Cortés Rodríguez, L. Sobre conectores expletivos y muletillas en el español hablado. Ágora. Málaga, 1991.
[15] Casado Velarde, M. Introducción a la gramática del texto del español. Arco/Libros. Madrid, 1993.
[16] Kovacci, O. El comentario gramatical. Teoría y práctica. Arco/Libros. Madrid, 1992.
[17] Lamíquiz, V. “Valores de *entonces* en el enunciado discursivo”. En Actas del III congreso Internacional del Español de América, II. Junta de Castilla y León. Valladolid, 1991.
[18] Briz Gómez, A. El español coloquial en la conversación. Esbozo de pragmagramática. Ariel. Barcelona, 1998.
[19] Calsamiglia Blancafort, H. y Tusón Valls, A. Las cosas del decir. Ariel, 1999.
[20] Alonso, L.; Castellón, I.; Padró, L. “Lexicón computacional de marcadores del discurso”. En Procesamiento del lenguaje natural Nº 29. Sociedad Española para el Procesamiento del Lenguaje Natural. España, 2002.
[21] Dale, R. y Knott, A. “Using linguistics phenomena to motivate a set of coherence relations”. En Discourse Processes 18(1)35-62. 1995.
[22] Knott, A. “A Data-Driven Methodology for Motivating a Set of Coherence Relations”. Tesis doctoral. Universidad de Edimburgo. Edimburgo, 1996.
[23] Marcu, D. “The rethorical parking of natural lenguage texts”. En ACL-97. Madrid, 1997.
[24] En la confección del archivo entradas, el equipo Infosur ha creado modelos para nombres, verbos, adjetivos y adverbios. En los modelos se agrupan las clases de palabras de acuerdo con la estructura morfológica. Así por ejemplo, los nombres que solo tienen forma masculina y forman el plural adicionando ‘s’ (‘ábaco’, ‘ábacos’), pertenecen al modelo 1; los que poseen forma masculina y femenina y también forman el plural agregando ‘s’ (‘abogado’, ‘abogada’, ‘abogados’, ‘abogadas’) pertenecen al modelo 3, etcétera.
[25] /marcd/estruc/orden/ap: Marcador Discursivo, Estructurador de la Información, Ordenador, de Apertura; /marcd/estruc/orden/cont: Marcador Discursivo, Estructurador de la Información, Ordenador, de Continuidad; /marcd/conect/aditiv: Marcador Discursivo, Conector, Aditivo; /marcd/conect/consec: Marcador Discursivo, Conector, Consecutivo; /marcd/refor/exp: Marcador Discursivo, Reformulador, Explicativo.
[26] 'EMS': Etiqueta Morfosintáctica, 'marcd': marcador discursivo, 'TMARCD': Tipo de Marcador Discursivo, 'refor': reformulador, 'CLREF': Clase de Reformulador, 'distan': distanciamiento; 'conect': conector, 'CLCON': Clase de Conector, 'contra': contraargumentativo, ‘exp’: explicativo; ‘recap’: recapitulativo.