

# Resolución de anáfora pronominal para el español usando el método de conocimiento limitado\*

Grigori Sidorov, Omar Olivas Zazueta  
*Laboratorio de Lenguaje Natural y Procesamiento de Texto,*  
*Centro de Investigación en Computación (CIC),*  
*Instituto Politécnico Nacional (IPN),*  
*Av. Juan de Dios Bátiz, s/n, Zacatenco, 07738, México D.F.*  
*sidorov@cic.ipn.mx*

## Resumen

*Se presenta una herramienta (un sistema) que permite hacer la resolución de anáfora pronominal para el idioma español implementando el método MARS con las modificaciones pertinentes. Una de las diferencias importantes es que en nuestro caso usamos la información sintáctica proveniente de un analizador sintáctico para la resolución de anáfora. Se hicieron experimentos que mostraron la precisión de 92.3% sobre un conjunto de pruebas pequeño.*

## 1. Introducción

El problema de la resolución de la anáfora ha sido una tarea importante en el campo de Procesamiento de Lenguaje Natural (PLN) durante años. Debido a su importancia este problema ha sido afrontado desde distintos puntos de vista en una variedad de sistemas de cómputo, por ejemplo, se puede mencionar la construcción de grafos conceptuales (Montes y Gómez *et al.*, 2001) como una de las aplicaciones más importantes de resolución anáfora. Existen dos tipos de métodos de resolución de anáfora: basados en conocimiento adicional de varios tipos y de conocimiento limitado (*knowledge poor*). Los trabajos realizados para el idioma inglés (Hobbs, 1978; Lappin and Leass, 1994; Kennedy and Boguraev, 1994; Baldwin, 1997; Mitkov, 1998) y para el idioma español (Fernández, 1998; Palomar *et al.* 2001), coinciden en la necesidad de utilizar la semántica como fuente esencial para la resolución adecuada y completa de la

anáfora, aunque todavía no hacen pleno uso de ella por falta de analizadores semánticos confiables. Se han planteado métodos de resolución enriquecidos que combinan la semántica y la sintaxis, lo hacen para el inglés, en dominios restringidos con definiciones puramente manuales de jerarquías y rasgos. De igual manera, otros métodos alternativos incorporan los papeles sintácticos en patrones de co-ocurrencia mediante estrategias puramente estadísticas (Dagan and Itai, 1991).

Sin embargo, los métodos más simples de conocimiento limitado se pueden aplicar en muchos casos, por ejemplo: textos de dominio restringido o de vocabulario controlado, con resultados bastante satisfactorios. La mayoría de los métodos de conocimiento limitado resuelven la anáfora sin realizar algún tipo de análisis lingüísticamente complejo, basándose en puras heurísticas o aplicando solamente algún tipo de análisis sintáctico parcial.

Mitkov y Stys (1997) proponen el sistema que necesita poca cantidad de conocimiento para la resolución de la anáfora pronominal en manuales técnicos tanto en inglés como en polaco. Utiliza la concordancia en número, género y persona como restricción y una serie de indicadores de antecedente (*antecedent indicators*) a modo de preferencias. Este sistema es una modificación del expuesto por Mitkov anteriormente escogiendo sólo un subconjunto de los indicadores de antecedente expuestos en ese trabajo (los que tras un estudio previo considera más adecuados para los manuales técnicos). Cada uno de estos indicadores asignará valores numéricos a los antecedentes, escogiéndose finalmente como

---

\* Este trabajo fue realizado con el apoyo parcial del gobierno de México (CONACYT, SNI) e Instituto Politécnico Nacional (SIP, COFAA, PIFI), México.

antecedente de la expresión anafórica el que tenga mayor suma de estos valores.

Mitkov y Stys trabajan sobre la salida de un etiquetador morfológico gracias al cual obtienen la información léxica y morfológica. No se usa ningún tipo de información semántica. El conocimiento que ellos utilizan se limita a una serie de reglas gramaticales correspondientes a sintagmas nominales, información morfológica (número, género y persona), una lista de términos y un conjunto de indicadores de antecedente (los cuales variarán en función del tipo de texto). En caso de empate se acuden a dos criterios para seleccionar el antecedente correcto: en primer lugar se escoge el antecedente con mayor valor para el indicador de reiteración léxica, y en segundo lugar en caso que todavía persista el empate, se escoge el más cercano.

No existe ningún método de resolución de anáfora para el español que no use mucha información semántica. En este artículo nuestro objetivo principal era implementar un método y construir una herramienta para resolución de anáfora pronominal para el español usando conocimiento limitado, es decir, no usar la información semántica, sin embargo, hacemos el uso de la información sintáctica.

## 2. El método implementado

### 2.1. Factores que se usan en el método

El método se basa en el cálculo de valores de varios factores.

**2.1.1. Lo definido (*definiteness*).** Los sustantivos “*definidos*” de las oraciones previas tienen más posibilidad de antecendencia en anáfora pronominal que los “*no definidos*” (los sustantivos definidos obtienen 0 y los *no definidos* son penalizados con -1). Se considera un sustantivo como *definido* si es modificado por un artículo definido o por un pronombre posesivo o demostrativo.

**2.1.2. Lo dado (*givenness*).** Los sustantivos de las oraciones previas que representan “la información dada” (tema)<sup>1</sup> se consideran buenos candidatos a antecedentes y obtienen una puntuación de 1. Los candidatos que no representan el tema obtienen 0.

**2.1.3. Verbos indicativos.** La evidencia empírica sugiere que debido a la relevancia de los sustantivos

que siguen a los verbos de la siguiente lista, esos verbos son buenos indicadores de los antecedentes. Si un verbo es miembro del siguiente conjunto: {*discutir, presentar, ilustrar, identificar, resumir, examinar, describir, definir, mostrar, verificar, desarrollar, revisar, reportar, enfatizar, considerar, investigar, explorar, determinar, analizar, sintetizar, estudiar, cubrir, evaluar, tratar*}, se considera que el primer sustantivo que lo sigue es un buen candidato para ser un antecedente (puntuaciones 1 para tal sustantivo y 0 para los demás).

**2.1.4. Reiteración léxica.** Los sustantivos léxicamente reiterados son buenos candidatos a antecedentes (una frase nominal obtiene una puntuación de 2 si en el mismo párrafo es repetida dos o más veces, 1 si es repetida una vez y 0 si no se repite). Los elementos léxicamente reiterados incluyen frases nominales sinónimas que a menudo pueden estar precedidas por artículos definidos o pronombres demostrativos. Además, una secuencia de frases nominales con el mismo núcleo es tomada en cuenta como reiteración léxica (por ejemplo, “*película para niños*”, “*película infantil*”, “*la película*”).

**2.1.5. Frases nominales no preposicionales.** Una frase nominal no preposicional pura tiene mayor preferencia que una frase nominal que forma parte de una frase preposicional (puntuaciones 0 y -1). Por ejemplo:

*Pon el libro<sub>i</sub> en el estante, asegúrate que los niños puedan alcanzarlo<sub>i</sub>.*

Aquí, “*el estante*” es penalizado (-1) por formar parte de la frase preposicional “*en el estante*”. Esta preferencia puede explicarse en términos de relevancia desde el punto de vista de la teoría del enfoque (*centering theory*).

**2.1.6. Preferencia de patrón de colocación.** Esta preferencia se les da a los candidatos que tienen un patrón de colocación idéntico a un pronombre (2 y 0). En caso de español, la preferencia de colocación está restringida a los patrones “frase nominal (pronombre), verbo” y “verbo, frase nominal (pronombre)”.

**2.1.7. Referencia inmediata.** En manuales técnicos, la pista de “la referencia inmediata” puede ser muy útil en la identificación del antecedente. La heurística usada es que en construcciones del tipo “... (*tú*)  $V_1$  NP ... *con* (*tú*)  $V_2$  lo/la (*con* (*tú*)  $V_3$  lo/la)”, donde *con* puede ser sustituido por cualquier preposición o conjunción {*y/o/antes de/después de* ...}, la frase nominal después de  $V_1$  tiene mucha posibilidad de ser el antecedente del

<sup>1</sup> Se utilizó la heurística de que la información dada es la primera frase nominal en una oración no imperativa.

```

V(SG,3PRS,MEAN) -> () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
N(SG,MASC) -> (obj) <*NCMS000> // coche (3) : coche \ *NCMS000
  N(SG,MASC) -> (comp) <*NCMS000> // nuevo (4) : nuevo \ *NCMS000
  ART(SG,MASC) -> (det) <*TIMS0> // un (2) : un \ *TIMS0
N(SG,MASC) -> (subj) <*NPMS000> // Pedro (0) : pedro \ *NPMS000
CONJ_C -> (coord_conj) <*CC00> // y (5) : y \ *CC00
  V(SG,3PRS,MEAN) -> (coord_conj) <*VMIP3S0> // usa (8) : usar \ *VMIP3S0
  PPR -> (obj) <*PP3MS00> // lo (7) : ello \ *PP3MS00
  ADV -> () <*RG000> // no (6) : no \ *RG000
$PERIOD -> () <*Fp> // . (9) : . \ *FP

```

**Figura 1. El árbol sintáctico.**

pronombre “lo/la” que sigue al  $V_2$ ; por lo tanto, se le da preferencia (2 y 0). Ejemplos:

*Para encender la impresora, presione el botón, y manténgalo; presionado por un momento.*

*Desengrape el papel, acomódelo, después cárguelo; en la bandeja.*

**2.1.8. Distancia referencial.** En oraciones complejas, las frases nominales de la sección anterior son los mejores candidatos a antecedentes de una anáfora, seguidos por las frases nominales de la oración anterior, después por sustantivos situados dos oraciones más atrás y finalmente por sustantivos de tres oraciones atrás (2, 1, 0, -1).

**2.1.9. Preferencia de términos.** Las frases nominales que representan términos del tema tienen más posibilidad de ser antecedentes en comparación con sustantivos que no lo son (1 y 0).

## 2.2. El algoritmo

El sistema desarrollado implementa una modificación del algoritmo MARS propuesto para el idioma inglés en (Mitkov, 1998) que será descrito a continuación. Posteriormente se mencionan las variaciones a este método que hice para adaptarlo para el idioma español. El algoritmo opera en cinco fases.

En la *fase 1*, el texto a procesar es parseado sintácticamente usando un parser que devuelve partes de oración, lemas morfológicos, funciones sintácticas, número gramatical y las relaciones de dependencia entre los elementos del texto que facilitan la extracción de frases nominales complejas. En la versión original del algoritmo no se usa la información sintáctica.

En la *fase 2*, se identifican los pronombres anafóricos. En nuestra implementación solo enfocamos en resolución de pronombres en tercera persona y posesivos en plural y singular que demuestren anáfora nominal de identidad de referencia.

En la *fase 3*, para cada pronombre identificado como anafórico, se extraen los antecedentes potenciales (candidatos) de la parte que encabeza la sección en la que aparece el pronombre y del texto precedente al pronombre en un límite de hasta tres oraciones o un párrafo, el que contenga la menor cantidad de texto. Una vez identificados, estos candidatos se sujetan a pruebas morfológicas y sintácticas. Se espera que los candidatos extraídos cumplan un número de restricciones para conformar el conjunto de candidatos contendientes, que son los que serán considerados más adelante. Inicialmente, se requiere que los candidatos concuerden con el pronombre respecto al género y número.

En la *fase 4*, se aplican factores preferenciales y factores represivos al conjunto de candidatos contendientes. Cada factor aplica una ganancia numérica a cada candidato.

En la *fase 5*, el candidato con el marcador más elevado es seleccionado como el antecedente del pronombre. Los empates son resueltos al seleccionar el candidato más reciente con el más alto marcador.

Las ganancias de los indicadores de antecedencia, como fueron propuestos en el método de Mitkov, fueron obtenidas en base a observaciones empíricas, llevando esta influencia de decisión a consideración, y nunca han sido consideradas como óptimas o exactas. Al cambiar las ganancias aplicadas por los indicadores de antecedencia, es posible obtener mejores tasas de éxito.

Dado que el marcador de un candidato contendiente es calculado al sumar la ganancia aplicada por cada uno de los indicadores, el algoritmo puede ser representado como una función con  $K$  parámetros, cada uno representando un indicador de antecedencia

$$score_k = \sum_{i=1}^{i=K} x_{k_i}$$

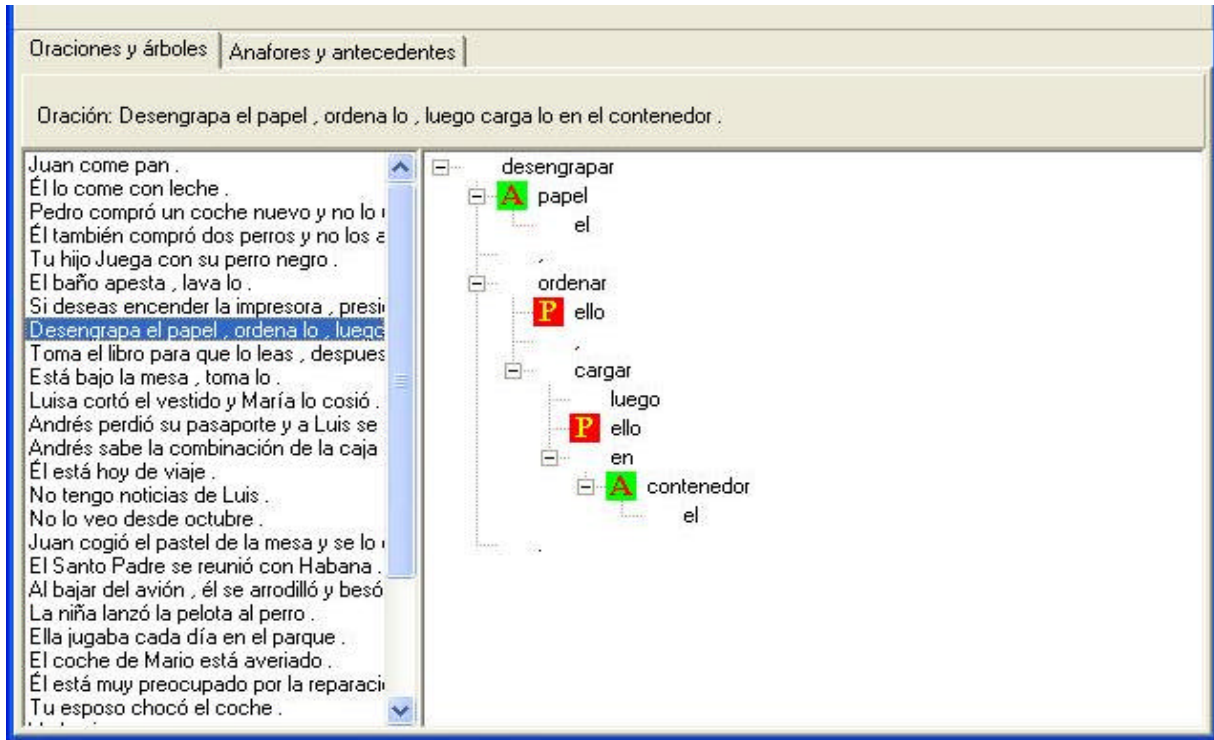


Figura 2. Representación de las oraciones.

donde  $score_k$  es el marcador compuesto asignado al candidato  $k$ , y  $X_{k_i}$  es la ganancia asignada al candidato  $k$  por el indicador  $i$ .

A continuación se mencionan las modificaciones al método original.

En la *fase 1*, para detección de frases nominales se utiliza el parser desarrollado en el Laboratorio de Lenguaje Natural del Centro de Investigación en Computación del Instituto Politécnico Nacional.

En la *fase 3*, es el usuario del sistema quien decide el número de oraciones de las que se extraerán los antecedentes potenciales, teniendo un límite de 10 unidades.

En la *fase 4*, se utilizaron los siguientes indicadores de antecedencia: lo definido (*definiteness*), lo dado (*givenness*), verbos indicativos, reiteración léxica, frases nominales no preposicionales, preferencia de patrón de colocación, distancia referencial, y además de estos indicadores, se incluyó un nuevo indicador, que además de verificar que cumpla el mismo patrón de colocación, verifica que se está utilizando el mismo verbo (2 y 0).

Otras diferencias están relacionadas con el manejo de género y las clíticas que existen en español a diferencia de inglés y carencia de necesidad de manejar el pronombre *it*.

### 3. Descripción del sistema

El sistema se basa en un analizador sintáctico que emplea una gramática extendida independiente del contexto con elementos de unificación. Este programa incorpora los resultados de la investigación para compilar patrones de manejo para verbos, adjetivos y sustantivos del español. Los resultados incorporados permiten clasificar las variantes generadas por el analizador de una forma cuantitativa, mediante los pesos asignados a las variantes de acuerdo a los valores de las combinaciones de subcategorización. Los pesos de las combinaciones de subcategorización son el resultado de un proceso de análisis sintáctico y de una extracción conforme a un modelo estadístico para determinar los complementos de verbos, adjetivos y algunos sustantivos del español a partir de un corpus de textos, véase, por ejemplo (Calvo and Gelbukh, 2006; Gelbukh *et al.*, 2005).

La entrada del sistema es una oración parseada, o, más bien, el árbol sintáctico correspondiente, por ejemplo, la frase *Pedro compró un coche nuevo y no lo usa* tiene el árbol que se presenta en Figura 1.

El sistema tiene dos partes principales. En la pestaña "Oraciones y árboles" se puede seleccionar cualquiera de las oraciones de la entrada y se muestra el árbol de

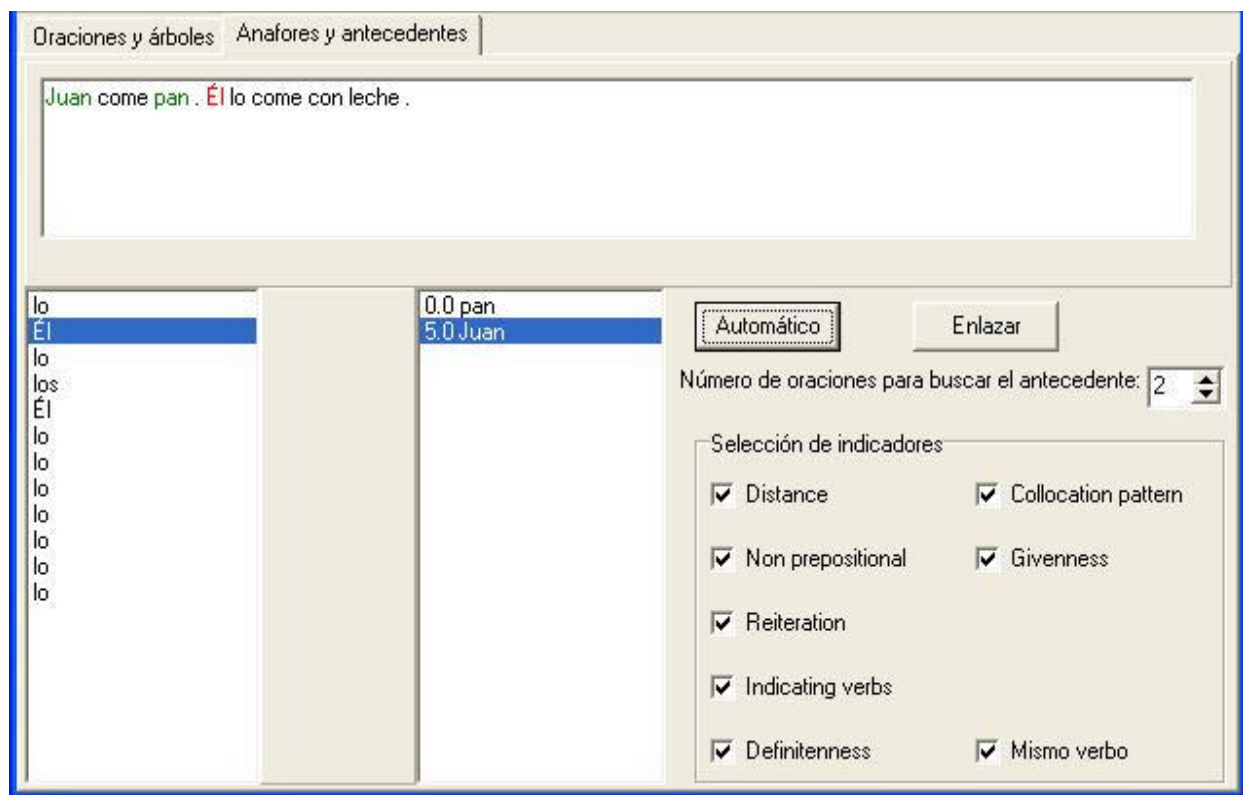


Figura 3. Ejemplo de funcionamiento del sistema.

dependencias correspondiente construido por el parser con posterior edición manual en caso necesario. Este árbol indica cuáles de sus elementos son pronombres (anafores) y cuales de ellos son frases nominales (candidatos a antecedentes), marcados como *P* y *A*, véase Figura 2.

En la pestaña “Anafores y antecedentes” se presenta la lista de anafores. Al seleccionar un elemento de esta lista aparecen las oraciones correspondientes en la parte de arriba, y los puntajes que alcanzaron los candidatos a antecedentes en la parte derecha. También se puede cambiar la lista de parámetros que usa el algoritmo.

Veamos un ejemplo de la resolución correcta del algoritmo, Figura 3.

*Juan come pan. Él lo come con leche.*

El anafora a solucionar es el elemento *Él* de la segunda oración, observándose que el sistema ofrece la solución correcta.

Ahora presentamos un ejemplo de un caso de la resolución incorrecta.

*El coche de Mario está averiado. Él está muy preocupado por la reparación.*

El anafora a solucionar es el elemento *Él* de la segunda oración. El algoritmo determina que el peso

de *coche* es 2.0 y de *Mario* es -1.0, dando una solución incorrecta.

#### 4. Evaluación

El método MARS original para el inglés se evaluó sobre un manual técnico en inglés de una fotocopiadora Minolta obteniendo una precisión del 95,8%, justificando los errores por falta de información sintáctica y semántica. Para el polaco también lo aplicaron obteniendo una precisión del 92,1%.

En nuestro caso para el español, de 26 anáforas detectadas, el sistema resolvió correctamente 24, obteniendo un 92.3% de precisión.

#### 5. Conclusiones

Se desarrolló la herramienta que permite hacer la resolución de anáfora pronominal para el español modificando el método de R. Mitkov, basado en conocimiento limitado. Otra diferencia importante es que en nuestro caso usamos el analizador sintáctico.

Se realizaron pruebas de funcionamiento del método usando textos en el español. El método mostró alta efectividad (92.3%) en los textos de prueba.

Como el trabajo futuro podemos mencionar lo siguiente:

- Hacer pruebas sobre una colección de textos más grande.
- Dar la posibilidad de evaluar las pruebas de manera automática.
- Comparar los resultados obtenidos con otros métodos de resolución de anáfora.

## 6. Referencias

- [1] Calvo, Hiram and Alexander Gelbukh. "DILUCT: An Open-Source Spanish Dependency Parser Based on Rules, Heuristics and Selectional Preferences." *Lecture Notes in Computer Science*, N 3999, Springer, 2006, pp. 164–175.
- [2] Dagan, Ido y Alon Itai "A statistical filter for resolving pronoun references." *Artificial Intelligence and Computer Vision*, 1991, pp. 125-135.
- [3] Gelbukh, Alexander, Hiram Calvo, Sulema Torres. "Transforming a Constituency Treebank into a Dependency Treebank." *Procesamiento de Lenguaje Natural*, No 35. Sociedad Española para el Procesamiento de Lenguaje Natural (SEPLN), 2005.
- [4] Hobbs, J. *Pronoun Resolution*. Research Report # 76-1, Department of Computer Sciences. City College, City University of New York, 1976.
- [5] Fernández, Antonio Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo

mediante gramáticas de unificación de huecos. Tesis Doctoral, Universidad de Alicante, España, 1998.

- [6] Kennedy, Christopher y Branimir Boguraev. "Anaphora for everyone: pronominal anaphora resolution without a parser." In: *Proceedings of 16th International Conference on Computational Linguistics*, vol. I, 1996, pp. 113-118.
- [7] Lappin, S., and M. McCord "Anaphora resolution in Slot Grammar." *Computational Linguistics*, 16, 4, 1990.
- [8] Mitkov, R., and M. Stys. "Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish." In: *Proceedings of the Recent Advances in Natural Language Resolution, RANLP, 1997*.
- [9] Mitkov, Ruslan. "Robust pronoun resolution with limited knowledge." In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*. Montreal, Canada, 1998.
- [10] Montes y Gómez, Manuel, Alexander Gelbukh, Aurelio López López, Ricardo Baeza-Yates. "Flexible Comparison of Conceptual Graphs." *Lecture Notes in Computer Science* N 2113, Springer-Verlag, 2001, pp.102-111.
- [11] Palomar, Manuel, Antonio Fernández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Zaiz-Noeda y Rafael Muñoz. "An Algorithm for Anaphora Resolution in Spanish Texts." *Computational Linguistics*, 27(4), 2001, pp. 545-567.