Instituto Politécnico Nacional

Centro de Investigación en Computación

Maestría en Ciencias de la Computación

Laboratorio de Lenguaje Natural y Procesamiento de Texto

Sistema de construcción de redes semánticas con detección de anáfora

TESIS QUE PRESENTA

Lic. Omar Alejandro Olivas Zazueta

PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

DIRECTOR DE TESIS

Dr. Grigori Sidorov

México, D. F., junio de 2006

ÍNDICE

ÍNDICE DE FIGURAS

RESUMEN

Para la construcción de redes semánticas y generalmente para cualquier procesamiento automático a nivel profundo, es necesario resolver la anáfora en textos.

Se desarrolló la herramienta (el sistema) que permite hacer esta resolución de anáfora pronominal para el español. Para eso se implementó el método de R. Mitkov para el español con las modificaciones pertinentes. También una de las diferencias importantes es que en nuestro caso usamos un analizador sintáctico para procesar los datos de entrada del método a diferencia del método original.

Se hicieron pruebas que dieron 92.30% de precisión sobre los datos de prueba que consistieron en 31 oraciones con 25 casos de anáfora. Para la evaluación se hizo comparación manual de los resultados del sistema con los datos obtenidos automáticamente por el algoritmo.

SUMMARY

For construction of semantic nets and, in general, for any kind of automatic processing at the profound level, it is necessary to have the possibility to resolve anaphora in texts.

In this thesis, I developed an instrument (a system) that allows performing anaphora resolution for Spanish language. I implemented the method of R. Mitkov with necessary modifications for Spanish. Also, one of the important changes was the application of Spanish syntactic analyzer for input data that adds more information for the method as compared to the original method.

I made experiments that give the precision of 92.30% for the test data that contain 31 phrases with 25 cases of anaphora. For the evaluation, I made manual comparison of the automatically obtained results and the correct output as determined by expert.

CAPÍTULO 1 INTRODUCCIÓN

En este capitulo describimos los objetivos de esta tesis así como la justificación y relevancia de la misma.

1.1 Objetivo general

Implementar el método y la herramienta (el sistema) para resolución de anáfora pronominal para el español para la construcción de redes semánticas.

Objetivos específicos

Para lograr el objetivo de este trabajo es necesario alcanzar los siguientes objetivos específicos:

- Desarrollar la interfaz entre el módulo de análisis automático morfológico y el sintáctico (*parser*).
- Adaptar el método de R. Mitkov de resolución de anáfora para el español. Igualmente hacer las modificaciones relacionadas con el uso de un analizador sintáctico automático.
- Implementar este método de resolución de la anáfora pronominal con las modificaciones pertinentes para el español.
- Integrar el método en un sistema de fácil acceso.
- Realizar pruebas de funcionamiento del método usando textos en español.
- Evaluar la efectividad del método desarrollado.

1.2 Importancia y Relevancia (Justificación)

El problema de la resolución de la anáfora ha sido durante años una preocupación en la comunidad de Procesamiento de Lenguaje Natural (PLN).

Esta tarea es considerada por muchos como una de las más importantes, la que ha sido afrontada desde distintos puntos de vista en una variedad de sistemas de cómputo. Algunos métodos de *conocimiento limitado* (*knowledge poor*), resuelven la anáfora sin realizar un análisis sintáctico, o realizando análisis parciales o completos. Los trabajos realizados para el inglés (Hobbs, 1978; Lappin y Leass, 1994; Kennedy y Boguraev, 1994; Baldwin, 1997; Mitkov, 1998) y para el español (Fernández, 1998; Palomar *et al.* 2001), coinciden en la necesidad de la semántica (aunque no hacen uso de ella) como fuente esencial para la adecuada resolución de la anáfora.

Debido a esta necesidad, diversos autores 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 e Itai, 1991).

No existe ningún método de resolución de anáfora para el español que no use mucha información semántica (no disponible todavía para el español); sin embargo, se basan en la información sintáctica.

En el Laboratorio de Lenguaje Natural del Centro de Investigación en Computación del Instituto Politécnico Nacional se cuenta con un parser sintáctico para el español que usamos para generar la entrada de nuestro algoritmo de resolución de anáfora.

1.3 Estructura de la tesis

La tesis se estructura como se detalla a continuación.

En el siguiente capítulo se mencionan los aspectos generales sobre el tema del discurso y la anáfora, los tipos de anáfora que pueden presentarse en función del marco en que sucede, en función del tipo de referencia y según el tipo de expresión anafórica.

En el capítulo 3 se presenta una revisión del estado del arte sobre las estrategias propuestas para el tratamiento del problema de la anáfora, las fuentes de información utilizadas para su solución y los algoritmos desarrollados para tal efecto.

El capítulo 4 trata sobre el sistema desarrollado para la solución de nuestro problema de estudio. Se describe inicialmente el método implementado, se hace una breve descripción del parser utilizado en el sistema y finalmente se hace una descripción del sistema.

En el capítulo 5 se muestran los experimentos realizados y la evaluación del sistema bajo dichos experimentos.

Se finaliza esta tesis con el capítulo 6 que muestra las conclusiones del trabajo realizado y las futuras direcciones de investigación. Al final se muestran las referencias bibliográficas que se emplearon en la realización del trabajo de tesis.

CAPÍTULO 2 Anáfora

El Procesamiento del Lenguaje Natural (PLN) es considerado una de las ramas principales de la Inteligencia Artificial (IA), que investiga y formula mecanismos computacionalmente efectivos que faciliten la interacción hombre máquina al estudiar una propiedad importante de la inteligencia humana: su capacidad de comunicarse por medio del lenguaje.

Dentro del campo de estudio del PLN, la resolución de la anáfora es uno de los problemas más difíciles pendientes de solución. Si se tiene la intención de construir máquinas que simulen inteligencia, éstas deben ser capaces de comunicarse con otros agentes en lenguaje natural, ya sean otras máquinas o personas.

Los sistemas de PLN intentan simular parte del comportamiento lingüístico humano; para lograrlo deben tomar conciencia de las estructuras propias del lenguaje, así como del conocimiento general acerca del universo del discurso. De esta forma, una persona participante en un diálogo sabe cómo hacer una combinación de palabras para formar una oración y posee un conocimiento del mundo en general que le permite participar en la conversación.

2.1 Anáfora y redes semánticas

Las redes semánticas son herramientas poderosas para diseñar mapas de conceptos, representándose en forma gráfica ideas y sus interrelaciones. Las redes semánticas o mapas conceptuales están compuestos por conceptos conectados por vínculos de relación. Supuestamente los mapas son el reflejo de la estructura mental que un sujeto tiene con relación a un tópico determinado. Estas relaciones pueden ser de carácter lógico (causalidad, identidad), representar el papel semántico que juegan unos nodos en relación a los otros (cercanía,

propietario, amigo) o simplemente representar una pertenencia tipológica, en cuyo caso el resultado suele denominarse jerarquía de tipos o taxonómica.

El proceso de construir redes semánticas exige la identificación de los conceptos más importantes de un contenido, siendo la experiencia un factor decisivo en esta dirección. El siguiente paso es el de conectar los conceptos que supone articular las posibles relaciones que existen entre ellos. El proceso de conexión continúa al tiempo que nuevos conceptos y relaciones aparecen.

Una etapa más importante en la construcción de las redes semánticas está relacionada con resolución de anáfora, porque es la única manera de tener una red semántica completa. Sin este conocimiento la red no se puede usarse en tareas de procesamiento automático. En el resto de la tesis nos centramos en la resolución de anáfora, entendiendo que los resultados de la resolución ya son partes de redes semánticas completas.

2.2 Discurso

Un discurso es una extensa secuencia de oraciones producidas por una o más personas con la intención de transferir o intercambiar información. Tal secuencia puede ser difícil de seguir: *cada oración debe ser entendida y agregada en un creciente banco de información*, y esto solo puede ser logrado si son claros los enlaces entre la oración actual y el discurso previo.

Por contexto del discurso se entiende *el conjunto de conocimientos y creencias compartidos por los interlocutores de un intercambio verbal y que son necesarios para producir e interpretar sus enunciados*.

Dentro del estudio del contexto del discurso se reconocen tres componentes: el sociocultural, el situacional y el lingüístico.

El contexto sociocultural es la configuración de datos que proceden de condicionamientos sociales y culturales sobre el comportamiento verbal y su

adecuación a diferentes circunstancias. Hay regulaciones sociales, por ejemplo, sobre cómo saludar o sobre qué tratamiento o registro lingüístico usar en cada tipo de situación.

El contexto situacional, es el conjunto de datos accesibles a los participantes de una conversación, que se encuentran en el entorno físico inmediato. Por ejemplo: para que el enunciado "*cierre la puerta, por favor*" tenga sentido, es necesario que haya ciertos requisitos o presuposiciones que son parte de la situación de habla: que haya una puerta en el lugar donde ocurre el diálogo, y que esté abierta.

El contexto lingüístico está formado por el material lingüístico que precede y sigue a un enunciado. En las actividades de lectura el contexto lingüístico es de gran importancia para inferir palabras o enunciados que no conocemos. Vale la pena puntualizar aquí una diferencia entre el texto oral y el escrito; en el texto escrito el contexto lingüístico incluye paulatinamente la información necesaria para construir el contexto situacional en su proceso de generación o interpretación; esto se debe a que, entre emisor y receptor, no es posible: la interacción de los sentidos con elementos del entorno común; ni la posibilidad de solicitar una corrección o una aclaración en el proceso de comunicación, como ocurre en una conversación oral.

2.3 Anáfora, referencia y correferencia

La etimología del término *anáfora* tiene su origen en la palabra αναφορα del griego antiguo, palabra compuesta por ανα (atrás, hacia atrás) y φορα (llevar) y denota el acto de hacer referencia a algo previamente mencionado.

Actualmente existen diferentes definiciones del término anáfora, pero en todas ellas subyace el mismo concepto y es la forma en que la anáfora nos permite hacer referencia a una determinada entidad[1], la cual es llamada antecedente. A la

[1] Cualquier persona u objeto que aparece implícita o explícitamente en el proceso de comunicación.

referencia abreviada se le llama *expresión o elemento anafórico*, o *anafor*. Veamos la definición de Hirst [**Error! Reference source not found.**]:

> *La anáfora es el mecanismo que nos permite hacer en un discurso una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de desabreviar la referencia y por consiguiente determinar la entidad a la que se alude.*

Franchini [**Error! Reference source not found.**] amplía esta definición en cuanto que permite diferenciar la anáfora respecto a la elipsis:

> *La pronominalización (un caso especial de anáfora) no es la sustitución por cero (caso de la elipsis), sino la sustitución por un elemento deíctico o proforma.*

Detalle evidenciado también en la definición de Covington [**Error! Reference source not found.**]:

> *La anáfora es el uso de palabras especiales que representan a individuos, situaciones u otras cosas ya referidas anteriormente.*

La definición hecha por Rico [**Error! Reference source not found.**] nos resalta otra característica importante, indicando que no es necesario que la entidad a la que se haga referencia esté mencionada explícitamente en el discurso:

> *La anáfora de manera general se define como la relación de referencia que se establece entre una forma lingüística y un objeto, una persona o una situación que ya han sido mencionados de manera implícita o explícita con anterioridad durante el proceso comunicativo.*

La posibilidad de que el antecedente no aparezca explícitamente en el texto, determina que la anáfora se desarrolla en el contexto denominado *situacional*. La anáfora se puede desarrollar en diversos contextos o marcos. Estos son el convencional, situacional o lingüístico:

- Para la anáfora desarrollada en el contexto situacional, los antecedentes son deducidos a partir de la situación concreta ya que no necesariamente aparecen explícitamente en el texto. Por ejemplo, en una situación concreta en la que una persona intenta tomar algún objeto y otra persona pronuncia la frase *no lo vayas a romper*, se tomará como antecedente al objeto que la primera persona intenta tomar, aunque este no haya sido mencionado explícitamente.

- La anáfora desarrollada en un contexto convencional es expresada mediante fórmulas lexicalizadas, fijas, con una pronominalización evidente en las que prácticamente resulta imposible identificar exactamente el elemento al que el pronombre hace referencia: *él la regó, averíguatelas tú solo* o *yo me la paso bien*.

- Dentro del contexto lingüístico, los antecedentes se encuentran explícitamente en el texto, como por ejemplo en *cuida al bebé y duérmelo*, donde la expresión anafórica denotada por el pronombre *lo* se refiere al sintagma nominal *el bebé*. En esta tesis, se trabajará exclusivamente con la anáfora que acontece en el contexto lingüístico.

Ersan y Akman [**Error! Reference source not found.**] hacen una distinción entre la anáfora que se da en el contexto lingüístico y la que ocurre dentro de el contexto situacional, llamándolas *referencias endóforas* y *referencias exóforas* respectivamente. Las referencias exóforas, denominadas sencillamente como exófora, hacen referencia a entidades ajenas al mensaje lingüístico. Las referencias endóforas, incluyen a la anáfora y la catáfora, donde el elemento referente hace alusión a una entidad dentro del mensaje lingüístico.

Ristad [**Error! Reference source not found.**] define el problema de la *resolución de la anáfora* como la forma de derreferir una expresión anafórica, y según él, la salida de un sistema que la resuelva debe ser una representación del conocimiento que se tiene del objeto al que hace referencia esa anáfora. La

referencia es un concepto central del lenguaje que ha sido ampliamente estudiado. Principalmente, el problema reside en determinar cómo las palabras son capaces de denotar conceptos y particularmente cómo una secuencia de palabras puede denotar un único concepto. Por ejemplo, una persona comenta a otra sin ningún propósito anterior: *El periodista entrevistó a Fox*. Lo más probable es que el receptor del mensaje, en el contexto adecuado sea capaz de determinar que *Fox* es el presidente de México y no cualquier otra persona llamada *Fox*. Y cuando posteriormente en el texto, utilicemos un pronombre para hacer referencia a esta persona *Fox*, realmente estaremos refiriendo al presidente del Gobierno de una manera indirecta: *él → Fox → Presidente de México*.

De esta manera, propiamente hablando, una expresión anafórica no hace referencia a su antecedente, sino al referente de la expresión que sirve de antecedente, por lo que se habla de correferencialidad entre expresión anafórica y antecedente. Es decir, ambos hacen referencia al mismo concepto: la expresión anafórica refiere indirectamente a través del antecedente, mientras que éste último refiere directamente. Para indicar *correferencialidad* se acostumbra indicar poniendo el mismo subíndice al antecedente y al elemento anafórico: *Mi madre$_i$ me saludó cuando ella$_i$ entró*. De esta forma, Ristad distingue entre *expresiones referentes* (*referring expressions*) y *elementos anafóricos* (*anaphoric elements*), basándose en el nivel de referencia de cada uno. Las expresiones referentes hacen una referencia directa ya que sus referencias no tienen como intermediarios a otros elementos lingüísticos. Mientras que los elementos anafóricos, por ejemplo los pronombres, mantienen una referencia indirecta, ya que se apoyarán en otros elementos lingüísticos (las anteriores expresiones referentes). Chomsky destaca en su teoría de *rección y ligamiento* (*Government and Binding Theory*) [**Error! Reference source not found.**, **Error! Reference source not found.** y **Error! Reference source not found.**], que los nombres propios están caracterizados porque no aceptan relación de correferencia con un antecedente: *Luis$_i$ dijo que Luis$_i$ vendría*. *Luis$_i$ insultó a Luis$_i$*. Esto se debe a su contenido referencial

completo. Chomsky nombra a esta clase de palabras como *expresiones referenciales*.

Esta idea de correferencialidad, permite hacer una distinción entre los conceptos de *antecedente* y *referente*, así como lo describen Brown y Yule [**Error! Reference source not found.**] y Rico [**Error! Reference source not found.**]. Según estos autores, el *referente* constituye la representación mental de los objetos evocados por el texto, mientras que el *antecedente* es la representación lingüística que estos toman en el mismo. En la presente tesis no se hará distinción entre estos conceptos, ni se utilizarán los términos *expresión referente* o *expresión referencial*, será llamado simplemente *antecedente*.

Hirst [**Error! Reference source not found.**] señala que una *expresión anafórica* y su *antecedente* son *correferentes* (*coreferential*), y al procedimiento de la determinación de el antecedente de una expresión anafórica se le llama *resolución*. Allen [**Error! Reference source not found.**], también hace uso de esta definición, ya que cuando un pronombre tiene cierto antecedente, en realidad se dice que ambos están haciendo referencia al mismo objeto, es decir, ambos correfieren (co-refer).

Ciertos lingüistas consideran que realmente el pronombre anafórico es solo un sustituto del sintagma nominal al que remite, por lo tanto el problema de la resolución de la anáfora consiste simplemente en una sustitución. No obstante, este enfoque es inadecuado en algunos casos como se muestra en la siguiente frase: *[El hombre que la$_i$ busca]$_j$ tendrá [la vida que él$_j$ desea]$_i$*, por lo cual si continuamos con el proceso de sustituciones, se observara que los pronombres no se terminaran.

En este punto de vista erróneo de la resolución de la anáfora como una sustitución es también remarcada por Convington [**Error! Reference source not found.**] mostrando que la anáfora es una *representación de entidades del discurso*, y no palabras o expresiones. Él mismo plantea los siguientes ejemplos en los que se

muestran los problemas a considerar en la resolución de la anáfora como una simple sustitución: *Si cualquier cliente no paga, muestra su balance*. Aquí indudablemente no sería correcto resolver esa anáfora como: *Si cualquier cliente no paga, muestra el balance de cualquier cliente*. Igualmente en el texto: *Muestra el primer registro. Ahora muestra el siguiente registro. Ahora bórralo.* En la última oración tampoco sería correcto resolver la anáfora como *Ahora borra el siguiente registro*, sino que la referencia sería al registro mostrado en la oración anterior.

Es necesario mencionar que existe la posibilidad de que la expresión anafórica y su antecedente no establezcan una relación de correferencia. Por lo tanto, en estos casos la expresión anafórica incluye un nuevo objeto el cual no esta explícitamente en el texto, y este se relaciona con otro objeto nombrado anteriormente. Allen [**Error! Reference source not found.**] denomina a este tipo de situaciones como una *anáfora superficial* (*surface anaphora*) y la define con aquella que hace una referencia parcial a parte de un antecedente.

Allen propone el siguiente ejemplo:

A. *Tell me [John's grade in CS]*$_1$.

B. *Give me [it in MT]*$_1$ *as well*.

C. *Give me [Mike's in GG]*$_1$ *too*.

Los objetos a los que hacen referencia las expresiones anafóricas *it in MT* y *Mike's in GG*, son referencias de objetos que no han sido nombrados anteriormente de forma completa, sin embargo, el contexto creado por la frase *A* permite interpretarlos de manera correcta.

En resumen, podemos concluir que la solución al problema de la anáfora puede ser una representación del conocimiento que se tiene del objeto al que se esta haciendo referencia en esa anáfora y no una sustitución de palabras solamente.

El problema de la anáfora, puede ser tratado desde diversas ciencias humanísticas, tal y como destaca Rico [**Error! Reference source not found.**] él menciona que el problema de la anáfora puede ser tratado por otras ciencias humanísticas. Por ejemplo: existen teorías de la anáfora en Sintaxis, Semántica, Pragmática, Psicolinguística, Filosofía del lenguaje, Lógica y Lingüística Computacional. Esto sucede debido a que los procesos que intervienen en el establecimiento de relaciones anafóricas abarcan un amplio espectro de las ciencias humanísticas:

- En Psicolinguística se realizan experimentos con los cuales se intenta descubrir la manera en la que las personas que intervienen una conversación pueden crear y comprender las referencias anafóricas dentro del proceso comunicativo.

- En Filosofía del Lenguaje se sugieren teorías que explican la manera en que una anáfora puede llegar a establecer una relación con el mundo real mediante una referencia a un antecedente.

- En Lógica se establecen gramáticas que tratan la anáfora entre otros problemas lingüísticos.

- La Lingüística Computacional tiene la finalidad de tratar la información lingüística a través de métodos informáticos, y de esta manera facilitar el tratamiento automático del lenguaje natural. Meya y Hubert[**Error! Reference source not found.**], mencionan que la Lingüística Computacional tiene como finalidad entender la manera en la que el ser humano se comunica, y de esta manera crear modelos y sistemas con los cuales se capaciten los ordenadores para que éstos puedan simular un comportamiento inteligente.

Para concluir esta sección en la que se ha tratado principalmente el problema de la anáfora, es importante mencionar que el objetivo primordial de esta tesis se

delimita al campo de estudio de la lingüística computacional, y que se pretende tratar con la anáfora que acontece en el contexto lingüístico.

2.3.1 Tipos de anáfora

En la actual literatura con referencia al tratamiento de la anáfora, se encuentran varios tipos: anáfora intraoracional, anáfora discursiva, anáfora superficial, anáfora profunda, etc. Estos distintos tipos de anáforas dan ciertas variaciones del concepto original y general de anáfora, por ejemplo en función del marco en que se trata la anáfora, ya sea éste la propia oración o todo el discurso, se hablará de la anáfora intraoracional y anáfora discursiva respectivamente. En función de la accesibilidad del antecedente: anáfora morfosintáctica, semántica y pragmática, de tal modo que la morfosintáctica será la que tenga una mayor accesibilidad, es decir, la que se derrefiere más fácilmente, y la pragmática la que tenga menor accesibilidad. También en función al tipo de referencia que se establezca entre la expresión anafórica y su antecedente se hablara de una anáfora superficial y anáfora profunda, en la que la profunda establece una relación de correferencia, mientras que la superficial no. Y finalmente podemos encontrarnos distintos tipos de expresiones anafóricas: pronombres, sintagmas nominales definidos, sintagmas nominales formados por el pronombre *uno* (*one*) y algunos modificadores, sintagmas nominales cuyo núcleo sea un adjetivo, adverbios y complementos circunstanciales. Cada una de estas expresiones anafóricas definirá un determinado tipo de anáfora que en ocasiones se denomina de una manera específica: anáfora pronominal, anáfora de tipo “*one*”, anáfora de tipo adjetivo, anáfora superficial numérica, anáfora verbal y referencias temporales o locales.

Estos matices permiten de alguna manera encuadrar la parte del fenómeno anáfora que se pretende resolver. En este caso haciendo un vínculo con la sección anterior, se trabajara con la anáfora que ocurre en el contexto lingüístico, y se elegirá de la variedad de anáforas existentes en la anáfora discursiva, es decir, se va a considerar para la resolución de todo el discurso y no únicamente la

oración en la que se encuentra la expresión anafórica. Nos concentraremos en la anáfora morfosintáctica, es decir, la que tiene mayor accesibilidad de su antecedente. Se tratará las referencias superficiales así como las profundas, también las relaciones de correferencia como las de no correferencia, y se intentará resolver la anáfora pronominal.

A continuación se mostrará detalladamente una clasificación de anáfora, la cual será dividida en cuatro subsecciones, cada una de ellas se corresponderá con cada uno de los siguientes criterios de clasificación: en función del marco en que sucede, en función de la accesibilidad del antecedente, según el tipo de referencia y según el tipo de expresión anafórica.

2.3.1.1 En función del marco en que sucede

Dependiendo del marco en el que sucede la anáfora, ya sea dentro de determinada oración o de un discurso, se hablará de una anáfora intraoracional o anáfora discursiva (interoracional).

ANÁFORA INTRAORACIONAL

La anáfora intraoracional sucede dentro de una oración, esto quiere decir que el antecedente y la expresión anafórica se localizan en la misma oración. El inconveniente de esta definición es que cuando se intenta emplear la resolución de la anáfora en componentes donde la oración no es un componente aislado, sino uno más dentro de una unidad mayor: el discurso. Por lo tanto, si se esta superando el marco de una oración, aquí se estará manejando otro tipo de anáfora la cual sería la anáfora discursiva.

ANÁFORA DISCURSIVA

Este tipo de anáfora es el contrario de la anáfora intraoracional, ya que la expresión anafórica y su antecedente no se localizaran en la misma oración. Aquí la anáfora actúa como un mecanismo de conexión textual, es decir, la anáfora actúa como un instrumento lingüístico el cual ayuda a mantener el discurso como

una unidad de sentido, debido a la creación de relaciones de unión entre las distintas partes del texto. De esta manera favorece a la conservación de un foco de interés anteriormente establecido mediante las entidades discursivas, las cuales crean unos puntos de atención en el discurso que son confirmados y consolidados gracias a las expresiones anafóricas.

En el actual trabajo se trabajará con la anáfora discursiva, debido a que es un fenómeno lingüístico básicamente discursivo, esto quiere decir que no se puede reducir solamente al marco de la oración. Por ejemplo: *Anoté la cantidad$_i$ que te adeudo en el cuaderno. Te la$_i$ daré luego por teléfono. Me saludó una chica$_j$. La$_j$ conocía, pero no recordaba su nombre.* En función de la accesibilidad del antecedente

Rico [**Error! Reference source not found.**], la anáfora puede ser clasificada de tres tipos de accesibilidad del antecedente, esto quiere decir que la facilidad con la que se puede derreferir (acceder al antecedente de una expresión anafórica concreta). Estos tres tipos son la anáfora morfosintáctica, semántica y pragmática, los cuales están resumidos en la Figura 1.

Una expresión anafórica típica con accesibilidad alta es el pronombre, y por otro lado las de baja accesibilidad son las anáforas semánticas y pragmáticas, debido a que son las expresiones que poseen mayor carga léxica y son las que aportan mayor información sobre el antecedente, permitiendo de esta manera un antecedente con menor accesibilidad. De esta forma los pronombres habrán de referirse a entidades muy accesibles, ya que de otra manera si se tiene poca información léxica difícilmente se resolverían, mientras que las anáforas semánticas podrán referirse a antecedentes con menor accesibilidad con ciertas garantías de éxito. Mollá [**Error! Reference source not found.**] específica aún más esta clasificación, él la divide principalmente en tres grupos en función de la accesibilidad: los nombres propios y sintagmas nominales definidos formarán parte del primer grupo, el segundo estará compuesto por los demostrativos, y el

tercero por los reflexivos. El primer grupo será el de menor accesibilidad y el tercero el de mayor.

Accesibilidad	***Anáfora***
Máxima accesibilidad	Anáfora morfosintáctica: • *Nominal*: referencia a un sintagma nominal. • *Verbal*: referencia a un verbo o SV. • *Oracional*
Accesibilidad media	Anáfora semántica: • *Sinonimia*. • *Hiperonimia*. • *Anáfora contextual*
Mínima accesibilidad	Anáfora pragmática

Figura 1. Tipos de anáfora en función de la accesibilidad del antecedente.

Mediante la aplicación de 4 parámetros básicos es determinado el grado de accesibilidad de un antecedente, como lo estudian Ariel [**Error! Reference source not found.**] y Mollá [**Error! Reference source not found.**]:

- Distancia, esta consiste en la cantidad de oraciones que existen a partir de la expresión anafórica hasta llegar a su antecedente (entre mas es la distancia es menos la accesibilidad).

- Prominencia, según esta los antecedentes que se localizan en posiciones temáticas de la frase tendrán una prominencia alta, por lo que serán más fácilmente accesibles.

- Competencia, aquí se encuentran una cantidad de entidades discursivas las cuales están en competencia por ser el antecedente de la expresión anafórica. Entre mas competencia haya habrá menos accesibilidad.

- Unidad: si en el mismo párrafo encontramos la expresión anafórica u su antecedente, aquí tendríamos un alto grado de accesibilidad.

En las siguientes tres secciones se estudiará a detalle los siguientes tipos de anáfora: morfosintáctica, semántica y pragmática.

ANÁFORA MORFOSINTÁCTICA: NOMINAL, VERBAL Y ORACIONAL

En la anáfora morfosintáctica las relaciones anafóricas se explican a través de criterios de criterios morfológicos y sintácticos, y son las que muestran mayor accesibilidad. Dentro de este tipo de anáfora, se especifican los siguientes tres subtipos: la anáfora nominal, anáfora verbal y la anáfora oracional, las cuales definimos a continuación:

- La anáfora nominal se entiende como aquella en la que se hace referencia a un sintagma nominal: *Tu esposo$_i$ choco el coche. Yo le$_i$ vi.*

- La anáfora verbal esta se construye principalmente en inglés, a través de la eliminación del verbo o *gapping*, y mediante el uso de un pro-verbo (como lo son el do o el have). En esta anáfora se puede referir a un verbo o a una frase verbal, como se muestra en el siguiente ejemplo: *María [cocino muy bien ayer]$_1$, pero Teresa [lo hizo]$_1$ muy mal.*

- La anáfora oracional esta se construye a través de una expresión anafórica en la cual el antecedente es toda una oración, como ocurriría en [*No deberíamos salir a cenar esta noche*]$_1$. *Yo no opino eso$_1$*.

En este trabajo se tratará específicamente la anáfora morfosintáctica, esto quiere decir que se dejará a un lado la anáfora semántica o pragmática, y mas específicamente en la anáfora nominal, es decir, se tratarán las expresiones anafóricas que tengan como antecedente un sintagma nominal.

ANÁFORA SEMÁNTICA: SINONIMIA, HIPERONIMIA, CONTEXTUAL Y FORZADA

En la anáfora semántica las relaciones anafóricas sólo pueden explicarse auxiliándonos de los criterios semánticos. Usualmente en esta anáfora, los

antecedentes no se encuentran claros en el discurso, y debido a esto los antecedentes son muy poco accesibles. En este tipo de anáfora, se pueden encontrar las siguientes variaciones:

- Sinonimia: mediante la sinonimia la construcción de las relaciones anafóricas es posible siempre y cuando existan dos o más equivalentes semánticos que pueden intercambiarse, y no confundan al lector. Un equivalente semántico funciona como una anáfora, mientras que el término que se sustituye es el antecedente. Por ejemplo, en la siguiente frase: *Pedro se quitó sus gafas$_i$… Él limpió sus lentes$_i$*, aparecen dos sinónimos: *gafas* y *lentes*, realizando el segundo de ellos la función de expresión anafórica, es decir, forma una referencia a la misma entidad nombrada previamente en el discurso: *las gafas de Pedro*.

- Hiperonimia: esta es cuando las relaciones anafóricas son establecidas a través de hiperónimos[2]: *No sabía que ese coche$_i$ es tuyo. Opino que es un buen automóvil$_i$.*

- Anáfora contextual: en esta se utiliza la expresión anafórica como un vínculo con una idea o un hecho que se deriva de una parte del texto anterior. Aquí el antecedente puede ser una única entidad, implícita o explícita, puede ser también toda una idea evocada por el texto, o surgir de una presuposición. En cualquier caso, para que se pueda interpretar la relación anafórica es necesario el conocimiento y la compresión del texto en su totalidad. Un buen ejemplo en el siguiente texto: *Él se limpió las gafas$_i$ y se las ajustó a la nariz. Su montura$_i$ y cristales$_i$ estaban húmedas.* La interpretación de la referencia anafórica de su *montura* y *cristales*, exige

[2] La hiperonimia es la relación de un término que abarca a otros semánticamente. Por ejemplo *semana* es el hiperónimo de *lunes, martes, miércoles, etc*., de la misma forma que *árbol* es el hiperónimo de *roble, pino, etc*. La relación inversa de la hiperonimia es la hiponimia que consiste en la relación de inclusión de un término en otro: por ejemplo *roble* es un hipónimo de *árbol*.

que el lector conozca las partes que componen unas *gafas*. Sólo si sabemos que las *gafas* tienen *montura* y *cristales* podemos inferir correctamente que los elementos que componen la expresión anafórica tienen como antecedente a *gafas.*

Hirst[**Error! Reference source not found.**] muestra un tipo de anáfora la cual podría ser incluida en la anáfora semántica: la anáfora forzada (*strained anaphora*). Esta anáfora tiene como particularidad que sus antecedentes deben ser léxicamente similares a palabras que ya se encuentran en el texto. Por ejemplo: *Pedro se convirtió en guitarrista porque pensaba que éste era un bonito instrumento*, la expresión anafórica éste se refiere a la guitarra, palabra léxicamente similar a la que aparece en el texto: guitarrista.

Para la resolución de la anáfora semántica se está aplicando actualmente la información almacenada en WordNet. Por ejemplo, Poesio, Vieira y Teufel [**Error! Reference source not found.**] muestran cómo aplican WordNet a la resolución de la anáfora. WordNet se enmarca dentro del conjunto de diccionarios y corpus electrónicos que aportan información diversa de naturaleza léxica, sintáctica y semántica.

ANÁFORA PRAGMÁTICA

Para finalizar tenemos la anáfora pragmática esta tiene una relación entre la expresión anafórica y su antecedente que es independiente los factores contextuales, por lo tanto el lector es capaz de identificar el antecedente de la anáfora independientemente del contexto. Esto es debido al contenido pragmático presupuestado por el discurso. De este modo, en el texto: *Esto es como en un libro: La isla del tesoro, La isla del coral, …* se hace imprescindible poseer el conocimiento pragmático que ayuda a interpretar los títulos La isla del tesoro y La isla del coral como libros. Este tipo de anáfora necesita para la localización de su antecedente que se construya una representación del mundo.

2.3.1.2 En función del tipo de referencia

Anteriormente se definió que el proceso de resolución de una anáfora, es aquél que consiste en derreferir la expresión anafórica para de esta manera determinar su antecedente, consiguiendo establecer entre ellos una relación de correferencia o no.

En función del tipo de relación que se establezca entre la expresión anafórica y su antecedente Allen [**Error! Reference source not found.**] nos distingue dos tipos de anáfora: anáfora profunda (existe relación de correferencia) y anáfora superficial (no existe correferencialidad).

Otros autores llaman a estos dos tipos de anáfora de manera diferente. Por ejemplo, Woods en [**Error! Reference source not found.** y **Error! Reference source not found.**] las denomina anáfora completa y anáfora parcial respectivamente. Y también han sido citadas en el trabajo de Hirst [**Error! Reference source not found.**], pero llamándolas en este caso anáfora con identidad de referencia (IRA, Identity of Reference Anaphora) y anáfora con identidad de sentido (ISA, Identity of Sense Anaphora).

ANÁFORA PROFUNDA

El concepto de anáfora profunda engloba la referencia completa a un objeto que ha aparecido previamente en el discurso, como podría ser la frase: *[El chico de pelo rojo]$_i$ compró un libro$_j$. Él$_i$ lo$_j$ compró en la tienda de Juan.* Es decir, la expresión anafórica está refiriendo el mismo objeto que el antecedente, o lo que es lo mismo, existe una relación de correferencia entre ambos.

ANÁFORA SUPERFICIAL

Allen [**Error! Reference source not found.**] define la anáfora superficial (*surface anaphora*) como aquella que realiza una referencia parcial a parte de un antecedente. Con ello, la expresión anafórica introduce un nuevo objeto que no aparece explícitamente en el texto y que se relaciona con otro objeto

anteriormente nombrado, es decir, no se establece una relación de correferencia entre la expresión anafórica y su antecedente.

Es de resaltar que en muchas ocasiones este tipo de anáfora viene acompañada de adverbios del tipo también o además que explícitamente avisan de la comparación con algo previamente nombrado en el texto.

La anáfora superficial se puede dar con diversos tipos de expresiones anafóricas:

- Por ejemplo, en la anáfora pronominal, donde las expresiones anafóricas están formadas por pronombres: *El hombre que gana su dinero$_1$ honradamente es mejor que el que lo$_1$ roba*, donde el pronombre *lo* referencia *al dinero* del segundo hombre y no al del primero. Partee [**Error! Reference source not found.**] destaca una clase de pronombres que se incluyen dentro de este tipo de anáfora superficial. A esta clase de pronombres los denomina perezosos (pronouns of laziness).

- En la anáfora verbal también se dan estos tipos de referencias parciales, por ejemplo: *Pedro [besó a su mujer]$_1$. Juan también [lo hizo]$_1$*, en la que al resolver la anáfora verbal, queda la interpretación de si *Juan besó a su mujer* o bien *besó a la mujer de Pedro*.

  Hardt [**Error! Reference source not found.**] denomina a este tipo de referencias identificación descuidada (*sloppy identity*). Este fenómeno ocurre cuando después de resolver la anáfora verbal o la pronominal queda una referencia intermedia sin resolver generalmente producida por adjetivos posesivos. Por ejemplo, en *Tomás quiere a su gato. Juan también*, después de resolverla nos quedaría: *Juan también [quiere a su gato]*, en la cual el gato referido podría ser el de Tomás o bien el de Juan.

- La anáfora de tipo “one” es un tipo especial de anáfora que se da en inglés en el que las expresiones anafóricas están formadas por el pronombre *one* junto con varios modificadores. Determinados casos de este tipo de anáfora

realizan este tipo de referencias parciales, por ejemplo la siguiente frase: *Peter bought [a blue pen]$_1$ yesterday. He has bought [another one]$_1$ today*, en la que another one introduce un nuevo objeto en el discurso.

- Casos en los que el antecedente aparece cuantificado. Estos casos aparecen en la literatura, por ejemplo en Ersan y Akman [**Error! Reference source not found.**], como anáforas que funcionan como variables sin ligar (*bound anaphora*), como por ejemplo las siguientes frases: *[Ningún niño]$_1$ admitirá que él$_1$ es somnoliento. [Ningún hombre]$_1$ se culparía a [sí mismo]$_1$.* Según se desprende de estos ejemplos es evidente que no existe correferencia entre la expresión anafórica y su antecedente, sino que únicamente se ha de ligar la variable correspondiente al sintagma nominal cuantificado (ningún niño) con la del pronombre (*él)*.

  En de este tipo de casos en los que el antecedente aparece cuantificado, se abre un nuevo campo de estudio tratado con asiduidad en la literatura actual, por ejemplo en Shalom y Nissim [**Error! Reference source not found.**] y Chierchia [**Error! Reference source not found.**], es el conocido como donkey pronouns o donkey sentences. El nombre de estos pronombres viene dado precisamente por el ejemplo del cual se parte: Ever*y man who owns a donkey beats it. Most farmers that have donkey beat it* o *If a man owns a donkey, he beats it*. O también con los siguientes ejemplos equivalentes: *Every person who has a credit card, will pay his bill with it* o *If a painter lives in a village, it is usually pretty*.

Para concluir este apartado, y siguiendo la reflexión hecha por Popowich [**Error! Reference source not found.**] para los pronombres reflexivos, podemos decir que tan importante es determinar el antecedente de una expresión anafórica, como el determinar la relación existente entre ambos. En el mencionado trabajo de Popowich, se muestran ejemplos en los que también para los pronombres reflexivos se pueden dar los dos tipos de anáfora: tanto profunda como superficial. Por ejemplo, en la frase *Sólo Juan se compadece a sí mismo*, al resolver la

anáfora tendríamos que *Sólo Juan compadece a Juan*. La frase original se puede interpretar como que *Juan es la única persona que se compadece a sí misma*, mientras que en la segunda sería que *Juan es la única persona que compadece a Juan*. La diferencia entre ambas interpretaciones la resuelve Popowich por medio de los índices que asigna a cada sintagma nominal al resolver la anáfora: en el caso que asigne el mismo índice se tratará de anáfora profunda, y en caso contrario será anáfora superficial.

En nuestro trabajo también afrontaremos estos tipos de referencias parciales producidas por la anáfora superficial. Es decir, se detectaran determinadas situaciones en las que no se establece una relación de correferencia entre la expresión anafórica y su antecedente dentro de la anáfora pronominal y de tipo adjetivo, tipos de anáfora que se tratarán con mayor detalle en la próxima subsección.

2.3.1.3 Según el tipo de expresión anafórica

Las expresiones anafóricas más habituales en lenguaje natural son los pronombres y sintagmas nominales definidos. Sin embargo, podemos encontrar otras expresiones. Por ejemplo, para el idioma inglés, tal y como resalta Covington [**Error! Reference source not found.**], podemos encontrar los llamados verbos anafóricos (los auxiliares como *do* o *will*) que se referirán a acciones o estados. Y también podemos encontrar los adverbios realizando referencias a tiempos y lugares. En esta subsección vamos a estudiar los siguientes tipos de expresiones anafóricas: pronombres, sintagmas nominales definidos, anáfora de tipo “one”, anáfora tipo adjetivo, anáfora superficial numérica, anáfora verbal, adverbios y complementos circunstanciales.

ANÁFORA PRONOMINAL

Las referencias originadas por los pronombres han sido tratadas profundamente en la literatura actual sobre la anáfora, pudiéndose producir tanto como anáfora intraoracional como anáfora discursiva. E igualmente pueden realizar tanto

referencias parciales como completas, tal y como ya hemos visto en la subsección anterior. También vimos cómo estas referencias tienen una alta accesibilidad, es decir, una distancia pequeña con su antecedente ya que presentan una carga léxica muy pequeña.

En las referencias surgidas por los pronombres destacan las creadas por los pronombres reflexivos, ya que como veremos posteriormente determinan cierta información especial que permite la identificación de su antecedente: han de tener su antecedente en la misma frase ocupando principalmente la función de sujeto: *Juan$_i$ se$_i$ peina*. Los reflexivos también se diferencian del resto de pronombres por la posibilidad de producir una lectura recíproca de la frase, tal y como ocurre en *[Angélica y Brenda]$_i$ se$_i$ despidieron*.

Además hay la probabilidad que un determinado pronombre no tenga antecedente. A este tipo de pronombres Hirst [**Error! Reference source not found.**] lo denominó pronombres no referenciales, a los que no considera como anáfora. Este es un fenómeno habitual en inglés principalmente con el pronombre *it* (se coloca en los casos de oraciones impersonales en las que no hay sujeto, ya que en inglés es obligatoria su presencia). Podemos encontrar varios ejemplos en las siguientes frases: *It is fortunate that Aisha will never read this book. It is half past two. It is raining.* Los casos equivalentes en castellano corresponderían a las oraciones impersonales en las que el sujeto semántico es desconocido, por lo que aunque esté omitido el sujeto pronominal, en realidad no constituyen una anáfora (no tienen antecedente): *Llaman a la puerta. Dicen que aumentará la carne. Se está bien aquí*. Para finalizar es importante señalar que un pronombre puede hacer referencia a toda una oración o situación, no únicamente a unos de los sintagmas nominales anteriores: *[El Rey recibió un atentado mientras iba en un vehículo]$_i$. Esto$_i$ causó terror en la ciudad*.

DESCRIPCIONES DEFINIDAS

Descripciones definidas estas son expresiones anafóricas compuestas por sintagmas nominales. Poesio, Vieira y Teufel [**Error! Reference source not found.**] basados en el estudio de una serie de artículos de Wall Street Journal, hacen una clasificación de las descripciones definidas. En dicha clasificación tenemos los siguientes tipos:

- Descripciones anafóricas: descripciones definidas que correfieren con sus antecedentes y comparten el mismo nombre principal. En el corpus de estudio contaron un 30% de descripciones definidas de este tipo.

- Referencias puente: aquellos usos de descripciones definidas basadas en el discurso previo que requieren algún razonamiento en la identificación de su antecedente textual (más que el simple emparejamiento de nombres idénticos). Se contaron un 20% de descripciones definidas de este tipo.

- Descripciones definidas basadas en el conocimiento común: aquellas que usan conocimiento presupuesto en el discurso, como podría ser el caso del sintagma nominal el gobierno. Se contaron un 47% de descripciones definidas de este tipo.

- Expresiones idiomáticas o casos dudosos: aquellas que no tienen un uso anafórico, como el sintagma nominal el pelo en la siguiente frase: *No me tomes el pelo*. Se contaron un 3%.

Allen [**Error! Reference source not found.**] explica las descripciones definidas como una referencia a un objeto el cual está normalmente determinado de forma única en el contexto, y elabora una importante distinción dependiendo si éstas son utilizadas de manera existencial o referencial:

- El uso existencial de una descripción definida por Allen pertenecería a aquel que afirma la existencia de un único objeto el cual satisface la descripción y que anteriormente no ha sido introducido en el discurso. Por ejemplo, en *La luna ha salido esta tarde a las 6:55* aquí esta haciendo

referencia al satélite natural de la tierra, y no a otras lunas a las que se podría hacer referencia en un contexto adecuado.

Covington [**Error! Reference source not found.**] propone que en la fórmula lógica se le debe añadir esta lectura existencial. Por ejemplo, para la frase *El Rey es calvo* propone la fórmula lógica[3]: *(∃ ! X) (rey (X) ∧ calvo (X)),* es decir, existe un único rey.

- El uso referencial utiliza la descripción para hacer referencia a un objeto anteriormente conocido. Es decir, en la lectura referencial de una descripción definida el oyente puede identificar el antecedente, y no solamente admitir que el referente existe (a diferencia de la lectura existencial) Por ejemplo, en la frase *el perro está enfermo* el oyente debe identificar qué el *perro* es del que se está hablando. Por ello, en esta lectura se debe identificar el antecedente o de otro modo se considerará que la frase está mal formada. Sin embargo la identificación del antecedente no se ha de entender como una detección absoluta como ocurriría en el siguiente texto: *Alejandro se compró un automóvil$_i$ ayer. El pagó mucho por ese automóvil$_i$*, aquí no es necesario saber con exactitud que automóvil es al que se está haciendo referencia, solo es suficiente el saber que se trata del automóvil que compró Alejandro.

En el siguiente ejemplo se destaca la importancia de estas descripciones definidas referenciales: *Un perro$_i$ ladró y un gato$_j$ maulló. Después el perro$_i$ persiguió al gato$_j$.* En la segunda frase se está haciendo referencia tanto al gato como al perro de la primera frase, en cambio si la frase hubiera sido de la siguiente forma: *Después un perro$_k$ persiguió a un gato$_m$*, esto no da a entender que no se trata del mismo perro ni del gato.

---

[3] Covington utiliza el cuantificador ∃ ! X para expresar: “existe un único X”.

Hirst [**Error! Reference source not found.**] distingue un caso particular dentro de las descripciones definidas con las referencias producidas por los epítetos: *Minnet$_i$ perdió su muñeca, [la triste niña]$_i$ está llorando*.

En función del tipo de expresión anafórica, podremos obtener información importante para su resolución, como podría ser la posible posición de su antecedente. Por ejemplo, sabremos que los pronombres tendrán su antecedente preferentemente en la misma oración o no muchas oraciones atrás[4], y dentro de los pronombres, los reflexivos lo han de tener en la misma oración. Sin embargo en las descripciones definidas puede existir una mayor distancia con su antecedente. Esto es así ya que éstas tienen mayor carga léxica que los pronombres, por lo que la referencia se puede encontrar aunque haya una mayor distancia entre antecedente y expresión anafórica.

ANÁFORA DE TIPO "ONE", ANÁFORA TIPO ADJETIVO Y ANÁFORA SUPERFICIAL NUMÉRICA

Dentro de la literatura actual el tratado amplio de un tipo de anáfora es conocido como de tipo "one" (one-anaphora). Su nombre viene de los casos producidos dentro del inglés en los que la expresión anafórica se forma mediante un sintagma nominal con la siguiente estructura: el pronombre *one* acompañado de diversos modificadores. Un ejemplo podría ser el siguiente: *Wendy didn't give either boy [a green tie-dyed T-shirt]$_1$, but she gave Sue [a red one]$_1$*, en la que el sintagma nominal *a red one*, introduciría la siguiente entidad *a red tie-dyed T-shirt*.

Aquí el inconveniente con este tipo de anáfora es que aparezca el pronombre *one* sólo (sin ningún modificador) y con un significado indefinido, es decir, sin referirse

[4] Se han llevado a cabo diversos estudios acerca del número máximo de oraciones anteriores a considerar en el proceso de búsqueda del antecedente de una anáfora pronominal. En este trabajo se han considerado dos oraciones anteriores tras un estudio previo del corpus sobre el que aplicamos nuestro sistema. Rico [143] estudió el corpus Susanne (en lengua inglesa) observando que incluso se encontraban antecedentes tres oraciones atrás.

a ningún antecedente, ya que no se trata de anáfora. Podemos encontrar los siguientes ejemplos: *Smoking gives one cancer. My boss makes one work hard. John makes one sick.* Y en castellano podemos encontrar también ejemplos equivalentes como en *Uno se siente incómodo*.

La frase del primer ejemplo en castellano sería: *Wendy no dio a ningún niño [una camisa verde de manga corta]$_1$, pero a Sue le dio [una roja]$_1$*. A este tipo de anáfora la denominamos de tipo adjetivo precisamente porque la detectamos por la presencia de un sintagma nominal en el que ha sido elidido el núcleo nominal, función que el adjetivo realiza temporalmente (hasta que se resuelva).

Dentro de la anáfora tipo adjetivo destaca la formada por las expresiones anafóricas el mismo, la misma y sus variantes en plural. Este tipo de expresiones anafóricas se refieren a sintagmas nominales muy recientes, normalmente justo el sintagma nominal que le precede: L*a gente que use la cafetera$_i$, debe limpiar la misma$_i$ después de utilizarla*.

Un problema que nos plantea la anáfora de tipo “one” y la anáfora tipo adjetivo, es que permiten tanto referencias parciales como completas tal y como ya vimos en la sección que trata sobre la anáfora pragmática. Por ejemplo, en *Compré [una pera verde]$_i$ y [otra roja]$_i$. Yo prefiero [la verde]$_i$*, en la anáfora de tipo adjetivo producida por otra roja realiza una referencia parcial (introduce un nuevo objeto en el discurso) mientras que en la verde realiza una referencia completa (es correferente con una pera verde). Un método para distinguir entre ambos tipos de referencia podría ser la siguiente regla simple propuesta por Hirst [**Error! Reference source not found.**]:

Se llevará a cabo una referencia parcial cuando venga acompañado de modificadores del sintagma nominal que lo diferencien del antecedente.

Es decir, en la siguiente frase: *Peter was watching [the car I bought]$_i$ and John’s car. He liked more [the one I bought]$_i$*, tendríamos una referencia parcial ya que aunque la expresión anafórica viene acompañada de modificadores, en realidad

tiene exactamente los mismos modificadores que el antecedente al que se refiere (tan sólo se ha sustituido el nombre, car, por el pronombre *one*).

Un tipo de anáfora similar a la anterior de tipo adjetivo es la denominada anáfora superficial numérica (*surface count anaphora*). En este tipo de anáfora las expresiones anafóricas presentan la siguiente estructura: determinante + adjetivo + modificadores. El componente adjetivo se especializa (para distinguirla de la anterior anáfora tipo adjetivo) en los ordinales, numerales y todas las expresiones anafóricas que signifiquen cierto grado de orden: el primero, el último, el tercero, el tres, etc., información que será especialmente útil para determinar su antecedente. Así podemos encontrar el siguiente ejemplo: *Ángel miró [al perro$_i$ y al gato$_j$], pero finalmente prefirió el primero$_i$*. Para solucionar este tipo de anáfora es necesario que además de tener el antecedente en la memoria del oyente, también se retenga el orden con el que aparecen los antecedentes dentro de la misma. Claro que esto sería ilógico cuando son mucho antecedentes porque el oyente no podría recordar el orden en que aparecieron los antecedentes en la oración, como se muestra en el siguiente ejemplo: *Esta mañana compré un libro, un bolígrafo, un lápiz, un sacapuntas, un diccionario y una libreta*. El cuarto que compré se rompió rápidamente. Este tipo de anáfora en inglés tiene una estructura similar a la anterior anáfora de tipo “one” ya que también incluyen como núcleo el pronombre one, por ejemplo: *the second one o the last one*, sin embargo es conveniente distinguirla ya que realmente introduce un nuevo tipo de información (la relación de orden) indispensable para su correcta resolución.

ANÁFORA VERBAL

Este tipo de anáfora es construida mediante la llamada eliminación del verbo o *gapping*, y la utilización de un pro-verbo (por ejemplo en inglés los auxiliares do o have). Esta anáfora puede referirse bien a un verbo bien a una frase verbal, como ocurre en la frase: *Peter danced with Jane at the party. John did it too*. También podemos encontrar la anáfora verbal en castellano habitualmente formada por un

pronombre y un auxiliar, como por ejemplo en *Pedro [jugó]$_1$ muy bien al tenis ayer, pero Juan [lo hizo]$_1$ muy mal*.

Este tipo de anáfora verbal tiene un cierto parecido con la ya citada elipsis verbal. La diferencia entre ambas es que en la elipsis se produce una sustitución por cero, mientras que en la anáfora existe la expresión anafórica con la que se realiza la referencia. El ejemplo anterior de anáfora verbal se podría conseguir de modo equivalente como la siguiente elipsis verbal: Pedro *jugó muy bien al tenis ayer, pero Juan ∅ muy mal*. Otra diferencia entre la anáfora y la elipsis verbal es referida por Franchini [**Error! Reference source not found.**]: la anáfora verbal no es posible cuando el verbo anteriormente mencionado en el enunciado ya lleva un complemento directo: *Juan bebió vino y María ~~[lo hizo]~~ ∅ agua mineral.*

ADVERBIOS Y COMPLEMENTOS CIRCUNSTANCIALES

Los adverbios y complementos circunstanciales igualmente pueden formar expresiones anafóricas tal y como ocurre en el siguiente ejemplo: *La iglesia está [detrás de la librería]$_i$. Angélica fue ahí$_i$ después del almuerzo*.

Hirst [**Error! Reference source not found.**] señala este tipo de expresiones anafóricas como referencias temporales o locales, mostrando que sus antecedentes consisten siempre en la localización temporal (o local) más reciente en el texto, tal y como lo muestra en el siguiente texto: *El despertador suena a [las 6 de la mañana]$_1$. [Las siguientes dos horas]$_1$ se pasan en tranquila meditación, y es entonces$_1$ cuando…*, donde la expresión las siguientes dos horas toma su antecedente de la anterior expresión de tiempo: a las 6 de la mañana. Y la última anáfora, entonces, tomará su antecedente a partir de la anterior, es decir, las siguientes dos horas.

CAPÍTULO 3 MÉTODOS DE RESOLUCIÓN DE ANÁFORA

En este capítulo se explicaran diversas estrategias que en la actualidad se están utilizando en el tratamiento y la resolución de la anáfora. Estas estrategias se dividen en dos tipos: estrategias integradas y estrategias alternativas. Las estrategias integradas se basan en el conocimiento, es decir, manejan una serie de fuentes de información que se consideran necesarias para el correcto tratamiento de la anáfora, mientras que las estrategias alternativas utilizan información estadística y principios de razonamiento con incertidumbre.

En las estrategias integradas se considera que en el proceso para establecer los posibles antecedentes de una determinada anáfora está regido por ciertas fuentes de información como son: sintáctica, semántica y pragmática. A continuación se mostrará un algoritmo genérico de este tipo de estrategia para el tratamiento de la anáfora:

Lo primero es identificar las expresiones anafóricas. Después buscar sus antecedentes utilizando diferentes fuentes de información.

En la actual literatura se encuentran distintos métodos de coordinación sobre estas fuentes de información. Todas ellas tienen en común un punto el cual es la distinción entre las restricciones y preferencias. Las restricciones tienden a ser absolutas, por lo tanto eliminan antecedentes factibles de una determinada anáfora. Por otro lado, las preferencias tienden a ser relativas por lo que requieren el uso de información adicional y tienen carácter consultivo. Las preferencias se suelen aplicar después de las restricciones, y sólo en el caso en que quede más de un candidato posible. Éstas tienen como objetivo el seleccionar uno de estos candidatos posibles. Una misma fuente de información puede ser tratada como restricción o preferencia en función del algoritmo o sistema en que se aplique para la resolución de la anáfora, o en función del tipo de expresión anafórica. Por ello,

en la próxima sección trataremos en profundidad cada una de estas fuentes de información independientemente unas de otras, tanto si se tratan como restricciones o como preferencias, dejando para la siguiente sección el estudio de los algoritmos que las utilizan, momento en el que se describirán diversos métodos de coordinación de estas fuentes de información.

3.1.1 Fuentes de información

Según hemos comentado en la introducción de este capítulo se necesitan diferentes fuentes de información para permitir un correcto tratamiento de la anáfora. En función de la información que se utilice también clasificamos los algoritmos que se están aplicando en la actualidad: integrados y alternativos. En esta sección expondremos cada uno de estos tipos de información en diferentes subsecciones: información morfológica, léxica, sintáctica, semántica, pragmática, sobre la expresión anafórica e información estadística y sobre tratamiento de corpus.

3.1.1.1 Información morfológica

La información morfológica necesaria para el tratamiento de la anáfora se corresponde con la concordancia en número, género y persona entre la expresión anafórica y su antecedente:

- En *Luis compró varios libros$_i$. Los$_i$ compró en la tienda de Carlos*, el pronombre los necesariamente ha de referirse a los *libros* y nunca a *Luis* al no concordar en número.

- Por ejemplo, en el siguiente texto: *Alejandro$_i$ y Minnet$_j$ son buenos amigos. Sin embargo ella$_j$ comenzó a evitarlo$_i$*, la concordancia en género constituiría suficiente información para determinar que el antecedente del pronombre ella ha de ser necesariamente *Minnet* y no *Alejandro*.

- En: *Juan y yo fuimos al cine* el pronombre yo al estar en primera persona se le asociará el antecedente del narrador de la historia, y nunca se le podría relacionar con Juan al estar en tercera persona.

Se ha mostrado la información que se ha utilizado tradicionalmente para la resolución de la anáfora, la cual es considerada usualmente como una restricción cuando se seleccionan los posibles antecedentes, tan sólo citar los algoritmos propuestos por Lappin y McCord [**Error! Reference source not found.** y **Error! Reference source not found.**]. No obstante los investigadores están consientes de que estas restricciones no siempre se cumplen. Tal y como lo detallan Rico [**Error! Reference source not found.**] y Hirst [**Error! Reference source not found.**] se destacan principalmente los siguientes casos:

1. Las expresiones anafóricas en plural puede tener un conjunto de antecedentes en singular, como por ejemplo en el siguiente texto: *[Mario y Diego]$_i$ fueron a Bolivia. Ellos$_i$ estuvieron allí durante cinco días*, en la que la anáfora formada por el pronombre en plural *ellos* tiene como antecedente un grupo de antecedentes singulares: *Mario y Diego*. Otro ejemplo más complejo que el anterior, sería cuando estos antecedentes singulares están separados, o sea, no aparecen coordinados en un sintagma nominal, como *en Juan$_i$ le dijo a Pedro$_j$ que ellos$_{i,j}$ deberían llamar ahora*. Este último ejemplo determina un caso especial de anáfora denominado antecedentes divididos (*split antecedents*) en el que se incrementa la dificultad de resolución al tener que decidir qué antecedentes se incluyen en este conjunto y cuáles no, tal y como muestran Williams, Harvey y Preston en [**Error! Reference source not found.**] sobre el siguiente texto:

   A move to stop Mr Gaitskell from nominating any more labour life peers is to be made at a meeting of labour MPs tomorrow. Mr Michael Foot has put down a resolution on the subject and he is to be backed

by Mr Will Griffiths MP for Manchester Exchange. Though they may gather some left-wing support…

en el que el pronombre *they* tiene como antecedentes a *Mr Michael Foot y Mr Will Griffiths*, pero no hay nada que impidiese también incluir a *Mr Gaitskell*, por lo que en estos casos se necesita información semántica y del dominio del texto para determinar las relaciones aquí descritas.

2. Y viceversa, es decir, las expresiones anafóricas en singular puede tener un antecedente en plural, como se muestra en el siguiente ejemplo: *En el zoológico, un mono correteaba entre dos elefantes$_i$. Uno$_i$ empujó al mono*, en el que la anáfora tiene como antecedente a uno de los elementos del conjunto referido por el sintagma nominal *dos elefantes*. Otro ejemplo sucedería cuando la expresión anafórica también está cuantificada por el cuantificador universal, en cuyo caso quedaría siempre con número singular, como en *Los niños$_i$ compraron un juguete. Cada niño$_i$ eligió el que quiso*.

3. Otro caso sucedería cuando se emplean nombres colectivos. Estos nombres pueden aceptar una referencia en singular o en plural, como sucedería en el texto: *The tribe$_i$ had been walking for two hours. (They$_i$ were / iti was) tired*, donde they designaría a los miembros del conjunto, mientras que *it* designaría al conjunto en su totalidad. Un ejemplo equivalente para el castellano ocurriría en: *El clero$_i$ es conservador. (A ellos$_i$ / a éste$_i$) no les gusta el cambio*, donde de nuevo, el pronombre plural (*ellos*) referiría a los miembros del conjunto, mientras que el pronombre singular (éste) referiría al conjunto en su totalidad.

4. Las expresiones anafóricas con género femenino puede tener un antecedente con género masculino, y el caso inverso. Un ejemplo

sucedería en el siguiente texto: W*ho is this Bresson$_i$? Is she$_i$ a woman*? en el que el pronombre *she* se refiere a una persona de nombre *Bresson* que probablemente estará marcada como mujer en la conciencia del oyente. Estos casos vienen determinados habitualmente por nombres que se pueden aplicar tanto a entidades del mundo con género masculino como femenino, y se selecciona uno de estos géneros precisamente por los modificadores que le acompañan.

Para trabajar estas excepciones de la concordancia en número, género y persona los diferentes algoritmos eligen especialmente dos soluciones: tratan por separado cada una de estas excepciones ya sea almacenándolas en el diccionario o en el mismo sistema de tratamiento de la anáfora; o no definen esta información como una restricción absoluta sino como preferencia relativa.

3.1.1.2 Información léxica

En la información léxica se puede incluir aquella que es referente al comportamiento de algunas palabras o grupos de las mismas en ciertas situaciones. Esta información comúnmente se almacena dentro del léxico o diccionario del sistema.

Grober y Beardsley [**Error! Reference source not found.**] plantean un ejemplo sobre la utilidad de esta información, a la que ellos denominan valencia implícita causal del verbo (*implicit verb causality o causal valence*), mostrando que la valencia causal de un verbo puede afectar a los antecedentes que han sido seleccionados de las expresiones anafóricas cercanas. De esta manera creen que en general las oraciones con la siguiente estructura SN_1 VERBO SN_2 porque {*él* / *ella*}… tienen una tendencia natural por la que el pronombre *él* o *ella* se refiera a SN_1 para determinados verbos (VERBO), a SN_2 para otros, y neutrales (igual SN_1 que SN_2) para otros. Por ejemplo, en *Tamara$_i$ ganó a Anahí porque ella$_i$ es una jugadora muy hábil*, el pronombre se referirá a SN_1, mientras que en *Diego golpeó*

a Vilmar$_i$ porque él$_i$ se había portado mal, donde el pronombre hace referencia a SN_2.

Normalmente esta información se trata como preferencia, ya que no es dificíl encontrar ejemplos en los que no se cumpla, por ejemplo: *Tamara$_i$ ganó a Anahí porque ella$_i$ es una jugadora muy torpe*, y en *Diego golpeó a Vilmar$_i$ porque él$_i$ es más fuerte*.

Mitkov y Stys [**Error! Reference source not found.**] utilizan información de cierto grupo de verbos ingleses para los que se prefiere como antecedente el primer sintagma nominal que los siga. Estos verbos son *discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal* o *cover*. Un ejemplo podría ser la siguiente frase: *This table shows a minimal configuration$_i$; it$_i$ does not leave much room for ….* También utilizan información procedente de los sustantivos, como en el caso de situaciones en las que el núcleo del sintagma nominal que precede al verbo es chapter, section o table, entonces se escogerá el antecedente que siga al verbo, o del mismo modo, en caso que el sintagma nominal está en la cabecera de la sección y forma parte de la oración donde está la expresión anafórica, entonces se escogería éste como antecedente.

Otra información léxica que se ha utilizado habitualmente en el tratamiento de la anáfora (por ejemplo también la utilizan Mitkov y Stys [**Error! Reference source not found.**]) es la reiteración del antecedente. Según esta información se preferirá el antecedente que aparezca repetidas veces en el texto, con mayor grado de preferencia cuantas más veces aparezca repetido.

3.1.1.3 Información sintáctica

La información sintáctica es muy útil para la resolución de la anáfora, es por ello que es muy utilizada. Esta información se adquiere mediante el análisis sintáctico del texto.

La información necesaria para realizar el análisis sintáctico es almacenada en lo que se denomina gramática.

Ya que se obtiene la información sintáctica del texto, se pueden formular distintas reglas con las que permitiría determinar o eliminar antecedentes de alguna expresión anafórica. Por ejemplo, la siguiente regla formulada por Hirst [46]:

Un pronombre que aparezca en la oración principal nunca puede hacer referencia a un sintagma nominal que aparece en una oración subordinada a esta.

Por ejemplo, en la siguiente frase: *Ella no fue a cenar porque Ana no la invitó*, aquí el pronombre *ella* jamás podrá referirse a *Ana*. En caso contrario, cuando el pronombre esta dentro de la oración subordinada, no necesariamente se cumple la regla. Por ejemplo, *Angélica llegó tarde al aeropuerto porque ella se quedó dormida*, se podría aceptar que *Angélica* y *ella* fuesen correferentes, aunque esto estaría correctamente escrito si se omite el sujeto pronominal: *Angélica$_i$ llegó tarde al aeropuerto porque $\varnothing_i$ se quedó dormida.*

Este tipo de información sintáctica se utiliza de modo especial en la resolución de la anáfora pronominal bajo la denominación de restricciones c-dominio o c-command (*constituent-command*). Por ejemplo, Reinhart [**Error! Reference source not found.**] define la relación c-dominio entre dos constituyentes o nodos de la estructura sintáctica que representa a la oración:

Para dos nodos A y B de una estructura sintáctica, A c-domina a B si y sólo si el primer nodo que domine[1] a A, también domina a B.

Y también define la relación por la que en un árbol o estructura sintáctica el nodo A domina al nodo B si y sólo si existe un camino en ese árbol desde A hasta B que

[1] El primer nodo que domina a A es el propio padre de A en el árbol que representa la estructura sintáctica.

los une en sentido descendente. En el mismo trabajo, Reinhart define las siguientes restricciones c-dominio:

1. Un sintagma nominal completo se debe interpretar como no correferencial con cualquier sintagma nominal completo al que c-domine.

2. Un pronombre debe interpretarse como no correferencial con cualquier sintagma nominal al que c-domine.

3. Un pronombre reflexivo debe interpretarse como correferencial únicamente con un sintagma nominal que lo c-domine dentro de un espacio sintáctico definido: su categoría mínima de gobierno (minimal governing category).

Se definirá como categoría mínima de gobierno de un nodo A, al menor nodo que determine la oración o un sintagma nominal que c-domine a A y a su nodo principal (el del constituyente al que pertenezca A, por ejemplo el nombre de un sintagma nominal, o el verbo de un sintagma verbal).

4. Un pronombre no reflexivo debe interpretarse como no correferencial con cualquier sintagma nominal que lo c-domine y que pertenezca a su categoría mínima de gobierno.

Podemos encontrar un ejemplo de la aplicación de la restricción 2 en la Figura 1. Aquí se puede observar cómo el pronombre SN_3 no c-domina a SN_1 ya que el primer nodo que domina a SN_3 es SP, el cual no domina a SN_1 por lo que esta restricción no impide que pueda considerarse como un antecedente posible. Sin embargo en la Figura 2, el primer nodo que domina a SN_1 es O, el cual sí que domina a SN_3, por lo que han de considerarse no correferentes. Igualmente, otro hipotético caso de catáfora sería *Él vio a Diego*, en el cual el pronombre nunca podría correferir con *Diego* ya que *él* lo c-domina.

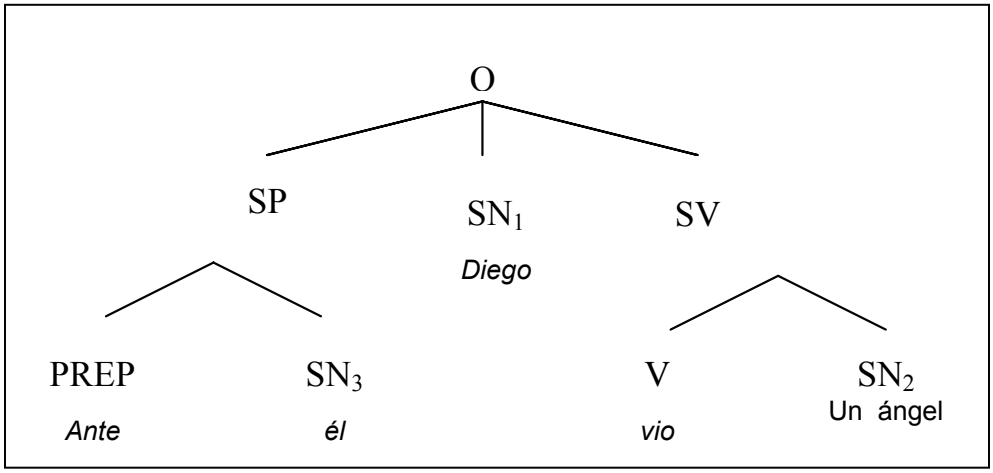


Figura 1. Ejemplo de aplicación de la restricción c-dominio número 2 en: *"Ante él, Diego vio un Angel"*.

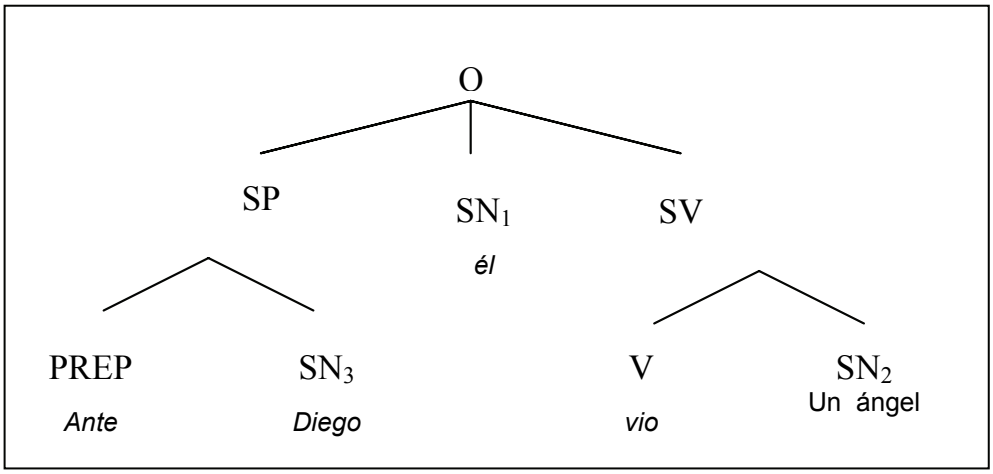


Figura 2. Ejemplo de aplicación de la restricción c-dominio número 2 en: *"Ante Diego, él vio un ángel"*.

Un ejemplo de la restricción 1 lo encontramos en la Figura 3, en la que podemos observar que tanto SN_3 como SN_1 son sintagmas nominales completos, y nunca podrán ser correferenciales ya que SN_1 c-domina a SN_3.

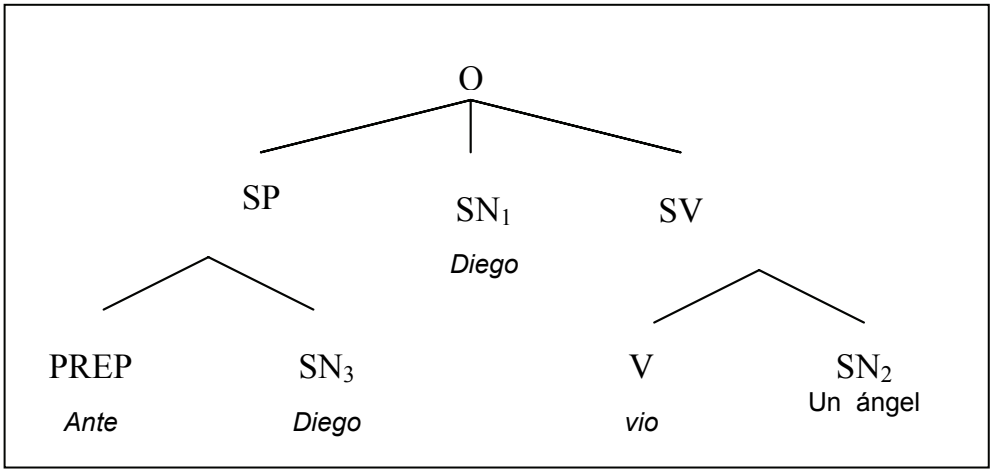


Figura 3. Ejemplo de aplicación de la restricción c-dominio número 1 en*: "Ante Diego, Diego vio un ángel"*.

La restricción 3 se ilustra mediante el ejemplo de la Figura 4: *Martha$_i$ se$_i$ molesta*. Esta restricción nos indica que un pronombre reflexivo debe ser interpretado como correferencial únicamente con un sintagma nominal que se encuentre en su categoría mínima de gobierno y que c-domine al pronombre. En esta figura SN_1 (*Martha*) c-domina a SN_2 (*se*) y a su vez pertenece a su categoría mínima de gobierno determinada por el nodo O, por lo que correfieren.

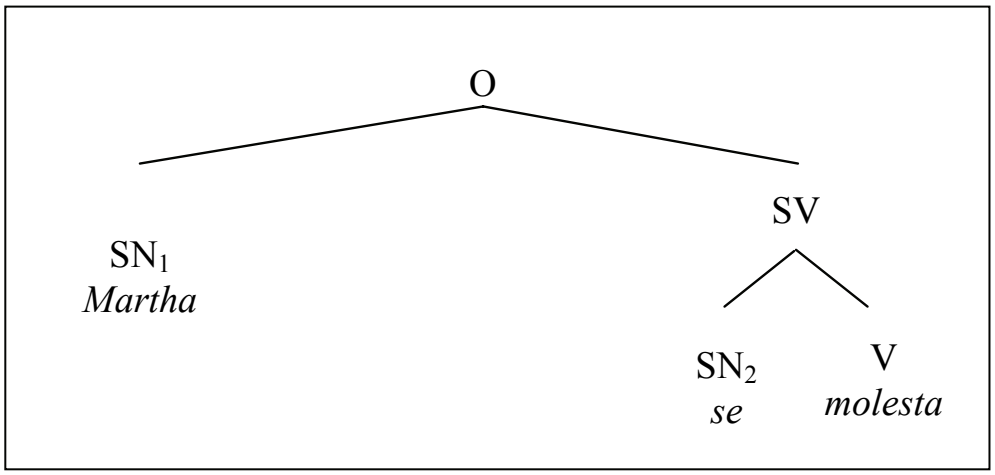


Figura 4. Aplicación de la restricción c-dominio número 3 en: *"Martha se molesta"*.

En la Figura 5 se muestra una frase similar a la anterior aunque se sustituye el pronombre reflexivo (*se*) por un pronombre personal (*le*): *Martha le molesta*. En este caso aunque no se incumple la restricción 2 (ya que le no domina a Martha), podremos rechazar el antecedente SN_1 gracias a la restricción 4 que nos exige que el antecedente no pertenezca a la categoría mínima de gobierno del pronombre (en este caso O). Encontramos otro ejemplo similar de la aplicación de esta restricción 4 en la frase *María llegó con ella*, en la que el pronombre no podría correferir con *María*, ya que aunque no c-domine a *María* (restricción 2), sí que se cumple la restricción 4: *María* pertenece a la categoría mínima de gobierno del pronombre. Igualmente se aplica esta restricción para evitar la correferencia en el *hijo*$_i$ *de él*$_j$, ya que el pronombre y el nombre *hijo* están dentro de la misma categoría mínima de gobierno que será el propio sintagma nominal que los incluye.

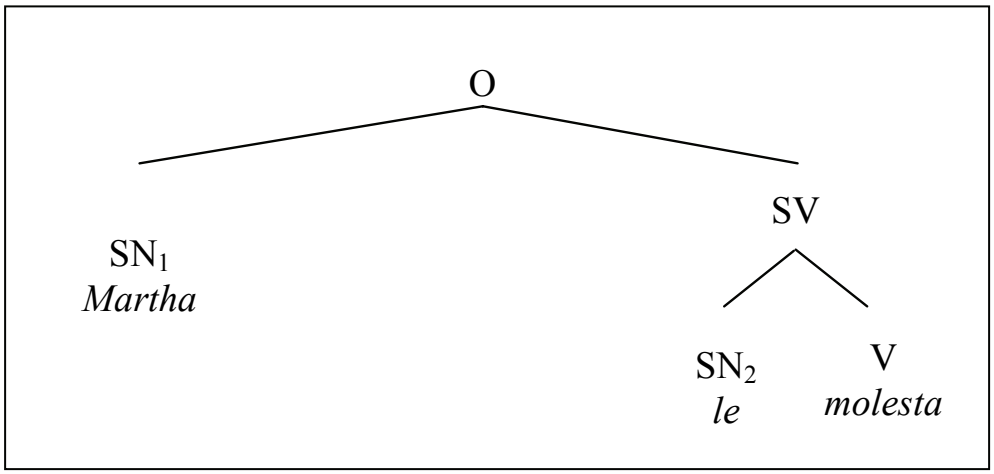


Figura 5. Aplicación de la restricción c-dominio número 4 en: *"Martha le molesta"*.

Finalmente, en el ejemplo de la Figura 6 se puede observar un caso en el que no se cumplen la restricción 2 ni la 4 (la categoría mínima de gobierno del pronombre *él* estará marcada por O3), por lo que el pronombre puede correferir con *Keller*.

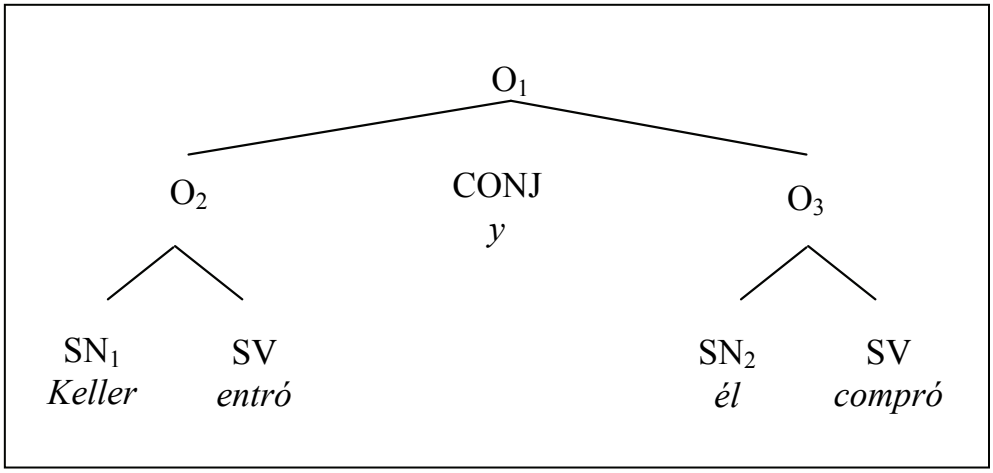


Figura 6. Ejemplo de correferencia en: *"Keller entró y él compró".*

Las restricciones c-dominio son utilizadas normalmente cuando se eliminan los antecedentes que no van a correferir con el pronombre en cuestión. Por esta razón se les llama *disjoint reference filters* (filtros que determinan la no correferencia) en el algoritmo propuesto en Kennedy y Boguraev [**Error! Reference source not found.**], o en el que enseguida se detalla propuesto por Lappin y McCord [**Error! Reference source not found.** y **Error! Reference source not found.**]. Las restricciones que se proponen en este último trabajo se basan en las siguientes definiciones que suponen una adaptación de la definición de c-dominio para las gramáticas de huecos (SG, Slot Grammars):

- Un sintagma P está en el dominio del argumento (argument domain) de otro sintagma N, si P y N son ambos argumentos de la misma palabra núcleo H. Por ejemplo: $Rocio_N$ $prefirió_H$ $éste_P$.

- P está en el dominio adjunto (adjunt domain) de N, si N es un argumento de una palabra núcleo H, P es el objeto de una preposición PREP, y PREP es un adjunto de H. Por ejemplo: $Martha_N$ $llegó_H$ con_{PREP} $ella_P$.

- P está en el dominio de sintagma nominal (noun phrase domain), si D es el determinante de un nombre N y, o bien P es un argumento de N (El_D $perro_N$

suyo$_P$) o bien P es el objeto de una preposición PREP y PREP es un adjunto de N (*El*$_D$ *hijo*$_N$ *de*$_{PREP}$ *él*$_P$).

Según Lappin y McCord [**Error! Reference source not found.** y **Error! Reference source not found.**] un pronombre P no es correferencial con un sintagma nominal N si alguna de las siguientes condiciones se cumple:

- P y N tienen características de concordancia incompatibles. Por ejemplo, en *La niña*$_i$ *mencionó que él*$_j$ *es hermoso*.

- P está en el dominio del argumento de N tal y como sucedería en la siguiente frase: *Aisha*$_i$ *prefirió esta*$_j$ o en la siguiente: *Armando*$_i$ *parece querer verle*$_j$.

- P está en el dominio adjunto de N. Ejemplos: *Aisha*$_i$ *llegó con ella*$_j$. *Who*$_i$ *did John say wants to sit near him*$_j$*?*.

- P está en el argumento de una palabra núcleo H, N no es un pronombre, y N está contenido en H. Ejemplos: *Éste es el niño*$_i$ *de quien ella*$_j$ *habló. Ésta es la niña*$_i$ *que él*$_j$ *conoció. A él*$_i$ *le encanta el perro de Juan*$_j$. *Who*$_i$ *did she*$_j$ *say Mary*$_k$ *kissed?.*

- P está en el dominio de sintagma nominal de N. Ejemplo: *[La mamá*$_i$ *de ella*$_j$*]*$_i$.

- P es el determinante de un nombre Q, y N está contenido en Q.

Las dos condiciones del final impiden la correferencia en las siguientes frases: *Su*$_i$ *retrato de Juan*$_j$ *es interesante. John*$_i$*'s portrait of him*$_j$ *is interesting. Su*$_i$ *descripción del retrato realizado por Juan*$_j$ *es interesante.*

Asimismo en los trabajos de Lappin y McCord, se pulen las anteriores reglas para establecer el siguiente algoritmo y con el resolver las expresiones anafóricas formadas por pronombres reflexivos y recíprocos (cada uno). Para ello define el concepto usado en la formulación del algoritmo de hueco argumento (o

complemento) mayor en base a la jerarquía de huecos argumento siguiente: *suj* > *agente* > *obj* > (*iobj* | *pobj*), donde *suj* es el hueco sujeto, *agente* es el agente de una oración pasiva, *obj* es el hueco objeto directo, *iobj* es el hueco objeto indirecto y *pobj* es el objeto de un complemento preposicional de un verbo. El algoritmo consitirá en la siguiente regla: un sintagma nominal N es un posible antecedente para un pronombre anafórico A si N y A no tienen incompatibilidades de concordancia y se cumple una de las siguientes cinco condiciones:

- A está en el dominio del argumento de N, y N llena un hueco argumento mayor que el que llena A. Por ejemplo: *They$_i$ wanted to see themselves$_i$. Aisha knows the people$_i$ who John introduced to [each other]$_i$*.

- A está en el dominio adjunto de N: *He$_i$ worked by himself$_i$. Which friends$_i$ plan to travel with [each other]$_i$*.

- A está en el dominio de sintagma nominal de N: *John liked Bill$_i$'s portrait of himself$_i$*.

- N es el argumento de un verbo V, hay un sintagma nominal Q en el dominio del argumento o en el dominio adjunto de N de manera que Q no lleva determinante, y A es:

    - un argumento de Q, o

    - A es un argumento de una preposición PREP y PREP es un adjunto de Q: *they$_i$ told stories about themselves$_i$*.

- A es el determinante de un nombre Q, y

    - Q está en el dominio del argumento de N y N llena un hueco argumento mayor que Q, o

    - Q está en el dominio adjunto de N: *[John and Mary]$_i$ like [each other]$_i$'s portraits*.

A continuación se da un ejemplo donde se muestra el efecto combinado de las dos últimas condiciones: *[John and Mary]$_i$ like [each other]$_i$'s portraits of themselves$_i$.*

Popowich [**Error! Reference source not found.**] su trabajo se trata ampliamente el fenómeno de las referencias originadas por los pronombres reflexivos, definiendo las siguientes restricciones sintácticas:

- Restricciones de localidad: en este tipo de restricción el pronombre y el antecedente aparecen en la misma oración, como por ejemplo en: *Minnet$_i$ se ama [ella misma]$_i$. Omar$_i$ encontró una fotografía de [el mismo]$_i$.* Esta restricción nos permite en la siguiente oración: *Ana dice que Luisa$_i$ esta enamorada de [ella misma]$_i$*, eliminar *Ana* como posible antecedente del pronombre.

- Restricciones c-dominio: el pronombre debe ser c-dominado por su antecedente. Esta restricción es similar a la anterior de localidad, por lo que exige que antecedente y pronombre se encuentren en la misma oración.

- Restricciones temáticas: el rol temático del pronombre no debe ser mayor jerárquicamente que el de su antecedente. Esta jerarquía de roles temáticos sería la siguiente: agente, localización, fuente, objetivo y tema, donde agente tendría la mayor jerarquía y tema la menor. De este modo se considerarían agramaticales las siguientes oraciones: *Luis$_i$ fue asesinado por [él mismo]$_i$*, ya que el pronombre toma el rol de agente, mientras que el sujeto toma el de tema. *Mario habló sobre Diego$_i$ a [él mismo]$_i$*, ya que el pronombre es el objetivo, mientras que Diego es el tema. Sin embargo una manera realmente correcta de decirlo seria: *Mario$_i$ habló a Diego sobre [él mismo]$_i$*.

Stuckardt [**Error! Reference source not found.**] ratifica estas restricciones de localidad proporcionando el siguiente ejemplo de correferencia: *The client$_i$ appreciates that the barber shaves him$_i$*, y cuya siguiente variación sería

incorrecta: *The client$_i$ appreciates that the barber shaves himself$_i$*. Aunque tal y como destaca Stuckardt, también se permiten referencias a elementos externos a ese dominio local, como ocurre en la siguiente frase: *The barber hears his$_i$ store about himself$_i$*.

De lo que se ha visto hasta ahora tenemos que, las restricciones sintácticas para la resolución de pronombres difieren en función del tipo de pronombre: ya sea reflexivo o no. Chomsky en su teoría de rección y ligamiento (Government and Binding Theory) [**Error! Reference source not found.**, **Error! Reference source not found.** y **Error! Reference source not found.**] lleva a cabo una comparación entre ambos tipos de pronombres. Donde nos muestra que los pronombres reflexivos y recíprocos necesariamente deben tener un antecedente para de esta manera poder interpretarlos, de lo contrario no seria posible. Por lo tanto son unidades faltas de referencias independientes: *Brenda$_i$ se$_i$ hizo daño [así misma]$_i$. [Pedro y Luis]$_i$ se$_i$ miraban [el uno al otro]$_i$*. Y es propio de este tipo de unidades (a sí mismo, el uno al otro) que su antecedente sea el sujeto de su misma oración. Esto también se cumple en las oraciones de infinitivo con formas reflexivas o recíprocas, ya que el sujeto léxicamente vacío del infinitivo es obligatoriamente correferencial con el sujeto de la oración principal: *Pedro$_i$ deseaba insultarse a [sí mismo]$_i$. [Pedro y Luis]$_i$ decidieron mirarse [el uno al otro]$_i$*. También se cumple para las oraciones subordinadas con sujeto vacío: *Pedro$_i$ aseguró que se había insultado a [sí mismo]$_i$. [Pedro y Luis]$_i$ decidieron que nunca más se mirarían [el uno al otro]$_i$*.

Según Chomsky hay una condición que es necesaria que se cumpla con el resto de los pronombres cuando tienen antecedente, la cual es que el antecedente se encuentre en una oración o cláusula distinta de la del pronombre (lo cual coincide con las restricciones c-dominio vistas hasta el momento): *Alejandro$_i$ jugó con él$_j$. Alejandro$_i$ lo$_j$ lavó. Alejandro$_i$ aseguró que él$_i$ no había ido al cine. Aisha$_i$ aseguró que ella$_j$ la$_i$ había perdonado*. Lo anterior se podría contradecir en la siguiente frase: *Angélica$_i$ se molestó por el regreso de Diego por ella$_i$*, seguramente la interpretación que se haría es la de no correferencia entre *ella* y *Angélica*, quizás

cabría la posibilidad de correferencia que se vería reforzada en el caso de utilizar el pronombre reflexivo ella misma. Por lo que respecta a su naturaleza semántica, el pronombre tiene cierta autonomía referencial ya que puede aparecer sin antecedente.

Como muestra Chomsky en los siguientes casos en los que el límite dentro del cual se desarrollan los fenómenos de ligamiento es el propio sintagma nominal: *Angélica criticó la devoción de Diego$_i$ por [sí mismo]$_i$. Angélica criticó la devoción de Diego$_i$ por él$_j$. Angélica$_i$ criticó la devoción de Pedro por ella$_i$. Angélica$_i$ criticó la devoción de Diego por [sí misma]$_i$.*

Pollard y Sag [**Error! Reference source not found.**] diferencian a su vez entre dos tipos de pronombres reflexivos: los reflexivos ordinarios y los reflexivos del tipo "foto" (picture noun reflexives, como el siguiente ejemplo: *John$_i$ found a picture of himself$_i$*). Pollard y Sag argumentan que ambos tipos de reflexivos no tienen el mismo tratamiento. Como ejemplos de reflexivos ordinarios muestran los siguientes[2]: *John$_i$ hates himself$_i$. The men$_i$ admired [each other]$_i$. Mary$_i$ explained Doris$_j$ to herself$_{i/j}$. Dana$_i$ talked to Gene$_j$ about himself$_{i/j}$. The men$_i$ introduced the women$_j$ to [each other]$_{i/j}$.* Y como ejemplos de reflexivos del tipo foto: *John found [a picture of himself]. The women selected [pictures of each other]. The men admired [each other's trophies]. The men introduced the women to [each other's spouses]*. En estos últimos Pollard y Sag indican que no requieren un antecedente que los c-domine tal y como ocurre en las siguientes frases: *A fear of himself$_i$ is John$_i$'s problem. The pictures of [each other]$_i$ with Ness made [Capone and Nitty]$_i$ somewhat nervous*.

Para concluir con el tema de las restricciones sintácticas para el tratamiento de la anáfora pronominal Hirst [**Error! Reference source not found.**] señala que se

---

[2] Indicaremos los casos en que exista ambigüedad en cuanto a los diferentes antecedentes para la misma expresión anafórica por medio del subíndice de cada antecedente separados por una barra (/). Por ejemplo: él$_{i/j}$.

necesita otro tipo e información sintáctica para el tratamiento de otros casos de anáfora: el paralelismo sintáctico entre constituyentes. Según Hirst, un sistema que intente resolver la anáfora es importante que tenga conocimiento acerca de paralelismo que particularmente importante para resolver la anáfora superficial numérica, ya explicada anteriormente (son aquellas que están formadas por ordinales y expresiones que signifiquen cierto grado de orden: el primero, el último, el tercero, etc.). de paralelismo que es particularmente importante para resolver la anáfora superficial numérica ya explicada anteriormente (aquellas que están formadas por ordinales y expresiones que signifiquen cierto grado de orden: el primero, el último, el tercero, etc.). En este tipo de anáfora se requiere almacenar la estructura superficial de la frase o al menos el orden de los posibles antecedentes. Por ejemplo, en la oración: *Minnet miró al perro$_i$, al gato$_j$ y al conejo$_k$, pero finalmente eligió el primero$_i$*, es preciso conocer el orden de los sintagmas coordinados (el perro será el sintagma nominal número uno, el gato el número dos, etc.) para poder resolver la anáfora el primero.

Mollá [**Error! Reference source not found.**] y Rico [**Error! Reference source not found.**] también hacen uso en sus trabajos de la noción del paralelismo sintáctico consistente en la preferencia de la expresión anafórica por aquellos candidatos que se encuentren en posiciones paralelas. Un ejemplo de lo anterior podría ser el siguiente: *Arturo$_i$ compró un pantalón$_j$. El$_i$ lo$_j$ estrenó en una fiesta*, en el que además se destaca cómo los roles semánticos de los pronombres reproducen los correspondientes roles de sus antecedentes. Carbonell y Brown [**Error! Reference source not found.**] nos señalan que las posiciones paralelas no solo pueden ser sintácticas, sino también pueden ser semánticas y pragmáticas, tal y como se verá en los próximos apartados en los que se estudiará a detalle estos nuevo tipos de información.

La estructura gramatical de los constituyentes analizados es otro tipo de información sintáctica interesante en el tratamiento de la anáfora, esta información es útil principalmente para resolver la anáfora de tipo “one” la cual consiste principalmente en los casos producidos en el inglés y su estructura es la

siguiente: el pronombre one acompañado de diversos modificadores. Por ejemplo, en *Wendy didn't give either boy [a green tie-dyed T-shirt]$_1$, but she gave Sue [a red one]$_1$*, el sintagma nominal a red one introduce la siguiente entidad: *a red tie-dyed T-shirt* para lo cual se está utilizando la información correspondiente a la estructura sintáctica del sintagma nominal antecedente de la expresión anafórica (a red one) para construir la nueva entidad del discurso a la que se está refiriendo (a red tie-dyed T-shirt).

3.1.1.4 Información semántica

La información semántica utilizada en el tratamiento de la anáfora se estructura habitualmente en forma de rasgos semánticos que se asignan a cada entidad del discurso. Esta información se suele utilizar obligando a que haya compatibilidad entre los rasgos semánticos de la expresión anafórica y su antecedente. Como ya hemos expuesto anteriormente, los pronombres disponen de poca carga léxica por lo que además tampoco dispondrán de estos rasgos semánticos. Estos rasgos semánticos los heredarán generalmente del verbo al que modifican. Por ejemplo, en la siguiente frase: *El ratón$_i$ se paró cerca del coche. Éste$_i$ encontró un trozo de queso$_j$ y se lo$_j$ comió*, se puede determinar que el pronombre *éste* se referirá al ratón y no al coche debido a que el verbo encontrar obliga a que el sujeto o agente de la acción sea del tipo ser vivo (*el ratón*), y no del tipo inorgánico (*el coche*). Estos rasgos semánticos se trasladan al pronombre ofreciendo una información muy valiosa que en determinadas situaciones nos permite determinar el antecedente correcto. Un caso similar ocurrirá con el pronombre lo. Este pronombre hace la función de objeto directo del verbo *comer*, por lo que hereda el rasgo semántico de ser algo comestible, rasgo que únicamente será compatible con el antecedente *queso*. Estas características de compatibilidad de rasgos semánticos se pueden determinar de diversas maneras. Por ejemplo, Hirst [**Error! Reference source not found.**] propone sencillamente tomar un valor de distancia semántica entre la expresión anafórica y sus posibles antecedentes para elegir el que tenga menor valor, es decir, esta distancia semántica se considerará únicamente como una métrica de qué grado de similaridad hay entre ambos.

Carbonell y Brown [**Error! Reference source not found.**] ellos nos hablan de utilizar este tipo de información como restricción, es decir, eliminarán todas aquellas entidades discursivas cuyos rasgos semánticos no concuerden con las restricciones semánticas exigidas por la expresión anafórica. Por ejemplo, en la frase: *Juan tomó el pastel$_i$ de la mesa y se lo$_i$ comió*, aquí seleccionaría el pastel en lugar de la mesa ya que el verbo comer impone en su objeto directo la restricción de una entidad comestible.

Carbonell y Brown en su trabajo también se habla de preferencias por posiciones semánticas paralelas, según las cuales las expresiones anafóricas tienden a seleccionar el antecedente que comparten la misma categoría semántica, aunque se manifiesten en diferentes posiciones sintácticas en la oración. Por ejemplo, en *Juan llevó la caja de Pedro a Luis$_i$. Él también le$_i$ llevó los libros de María*. Aquí el pronombre le y el sintagma nominal Luis comparten el rasgo semántico de receptor.

Rico [**Error! Reference source not found.**] observa que siempre debe existir consistencia semántica en la relación anafórica, y define tres tipos de consistencia semántica:

- La consistencia entre la expresión anafórica y su antecedente que ocurre mediante el empleo de sinónimos, hiperónimos[3], relaciones del todo a una parte, del grupo al individuo y de hipónimos[4]. Este tipo de consistencia semántica, se manifiesta normalmente en aquellas expresiones anafóricas que tienen mayor carga léxica (descripciones definidas).

---

[3] Se dice que un objeto es hiperónimo de otro si lo abarca semánticamente, es decir, es más general. Por ejemplo semana es el hiperónimo de lunes, martes, miércoles, etc., de la misma forma que árbol es el hiperónimo de roble, pino, etc.

[4] Se dice que un objeto es hipónimo de otro si éste es una especialización del segundo. Por ejemplo roble es un hipónimo de árbol.

- La consistencia entre el verbo y la expresión anafórica. La expresión anafórica debe obedecer siempre las restricciones semánticas impuestas por el verbo. Por ejemplo, en: *The jury recommended that Fulton legislators act to have these laws$_i$ studied and revised to the end of modernizing and improving them$_i$*. El pronombre *them* refiere a *these laws* gracias a la necesidad de mantener la consistencia semántica con los verbos *improve* y *modernize*. Por el contrario, si en lugar de estos dos verbos, introducimos el verbo *fulfil* (c*umplir*), *them* se referirá a *Fulton legislators*.

- La consistencia dentro de la oración en la que se encuentra la expresión anafórica. Esta es la más compleja porque utiliza el conocimiento del mundo para poder identificar el antecedente. En la oración: *The jury said it found the court has incorporated into its operating procedures*…, la expresión anafórica its operating procedures (procedimientos operativos) se refiere a the court (el tribunal), ya que the jury (el jurado) no tiene *operating procedures*. Aunque, como esta información es la más compleja de obtener, Rico encontró en su análisis del corpus Susanne que las relaciones anafóricas en las que el criterio semántico de inferencia es el definitivo para identificar el antecedente son muy poco numerosas (alrededor de un 6% del total de expresiones analizadas).

3.1.1.5 Información pragmática

Dentro de esta subsección comentaremos la información pragmática dividida en dos tipos que compondrán los dos siguientes apartados: información del mundo exterior e inferencia, y la información referente a la construcción del discurso.

INFORMACIÓN DEL MUNDO EXTERIOR E INFERENCIA

La información pragmática es fundamental en algunas ocasiones en las que hay cierta ambigüedad a la hora de elegir entre varios antecedentes, por ejemplo en la oración: *Ángel$_i$ le dijo a Rodrigo$_j$ que él$_j$ era el hombre más gordo que había conocido*, en este ejemplo tenemos claro quien es el hombre más gordo, Ángel o

Rodrigo, ya que empleando el conocimiento del mundo exterior deducimos que la autocrítica no es muy común, como lo es la crítica ajena. La información pragmática también nos permite introducir cierta información del mundo exterior e inferencia de conocimiento aparte del que nos traen las propias palabras del lenguaje y sus usos.

Otro ejemplo sería el que se presenta en la siguiente frase: *No les des plátanos$_i$ a los monos$_j$ porque estos$_i$ están verdes*, en la que se puede concluir que el pronombre *estos* se referirá a los *plátanos* porque sólo ellos y no los monos tienen la capacidad de estar verdes. O la siguiente: *Si [una bomba incendiaria]$_i$ cae cerca tuyo, no pierdas [la cabeza]$_j$. Ponla$_i$ en una cubeta y cúbrela$_i$ de arena*, en la que es necesario saber que el objeto peligroso en este contexto es la bomba y no la cabeza. O en la frase muy utilizada en los textos sobre anáfora: *If the baby$_i$ does not thrive on raw milk$_j$, boil it$_j$*, donde se ha de conocer que la leche puede ser hervida, mientras que los bebés no.

Mollá [**Error! Reference source not found.**] nos propone el siguiente ejemplo en el que nos muestra la importancia de la información del mundo exterior. En el ejemplo se presentan dos frases que solo difieren en una palabra, aunque esta palabra es suficiente como para que cambie el antecedente del pronombre *él*:

- *Anahi$_i$ regaló a Rocío su perro porque ella$_i$ lo odiaba*.
- *Anahi$_i$ regaló a Rocío su perro porque ella$_i$ lo necesitaba*.

Carbonell y Brown [**Error! Reference source not found.**] para ellos las restricciones pragmáticas son las restricciones impuestas por la acción que se está realizando en la oración. Por ejemplo, *Diego le dio a Carlos una pera. Él se comió la pera*, aquí el verbo dar está imponiendo la condición que la persona que da algo a alguien deja de tener ese algo. En este caso si el antecedente fuera Diego y no Carlos se estuviera estableciendo un conflicto entre la condición del verbo dar y la de comer que establece que para que alguien se pueda comer algo es preciso que ese alguien tenga ese algo.

INFORMACIÓN REFERENTE A LA CONSTRUCCIÓN DEL DISCURSO

En la información pragmática también se incluye cierta información referente a la construcción del discurso en el que se desarrolla la anáfora. Por ejemplo Carbonell y Brown [**Error! Reference source not found.**] consideran como un criterio de preferencia pragmática según el cual establece que debe darse mayor prioridad a los posibles antecedentes que se encuentren en posiciones temáticas o topicalizadas. Como por ejemplo *Rocío dijo a Brenda de ir a México. Por qué hizo ella esto?*, en donde Rocío se encuentra en una posición temática, por lo cual sería el antecedente preferido por el pronombre ella. Igualmente ellos consideran que son preferibles a igualdad de condiciones los antecedentes que se encuentren en la misma oración que la expresión anafórica y además seleccionan como antecedente preferido aquel que ayude a mantener la cohesión discursiva. Por ejemplo *Raúl viajó de España a México. Ángel fue allí también*, en la primera parte del discurso realiza una acción de movimiento de España a México. Y en la segunda oración se debe interpretar el allí como una referencia de *México* si se desea mantener la cohesión discursiva.

Este tipo de información es esencial en los algoritmos que se basan en la teoría del foco del discurso, teoría que plantea un tratamiento especial para la anáfora a partir del estudio de la estructura del discurso. Estos algoritmos se basan en considerar como primer antecedente posible el foco del discurso, conceptos que presentaremos con todo detalle mas adelante.

Rico [85] realizó un trabajo muy interesante, en el que estudió sobre un corpus de frases en inglés la relación de diversos aspectos del discurso con el tipo de expresión anafórica. Los aspectos estudiados fueron la distancia, competencia, prominencia y unidad. Los resultados de este estudio los mostramos en los próximos apartados.

Distancia

Los resultados del estudio referido en cuanto a la distancia entre la expresión anafórica y su antecedente se encuentran en la Figura 7. Aquí es considerada la distancia como el número de frases existentes entre la expresión anafórica y su antecedente. Según se puede apreciar en estos resultados, las descripciones definidas y descripciones con demostrativo permiten mayor distancia con su antecedente. Esto es así debido a que son precisamente estas expresiones anafóricas las que tienen mayor carga léxica, esto es, comparten más rasgos con su antecedente por lo que éste se identifica con mayor facilidad. Los pronombres permiten generalmente muy poca distancia con su antecedente. Sólo lo permiten (16,3%) cuando entre la expresión anafórica y el antecedente no se encuentra ningún otro antecedente que pueda competir.

Distancia	***Expresión anafórica***			
	Descripción definida	***Descripción demostrativo***	***Pronombre personal***	***pronombre posesivo***
0	**6.35%**	**4.4%**	**30.8%**	**63.1%**
1	**39.4%**	**75.5%**	**46.8%**	**27.8%**
2	**10.0%**	**13.3%**	**5.9%**	**3.2%**
3	**44.1%**	**6.6%**	**16.3%**	**5.7%**
TOTAL	**99.8%**	**99.8%**	**99.8%**	**99.8%**
0= El antecedente se encuentra en la misma oración **1= El antecedente se encuentra una oración antes** **2= El antecedente se encuentra dos oraciones antes** **3= El antecedente se encuentra mas de dos oraciones antes**				

Figura 7. Variación de la distancia en el tipo de expresión anafórica.

Los anteriores resultados explican el criterio de preferencia normalmente utilizado en innumerables trabajos acerca de la anáfora, por ejemplo Brown y Yule [**Error! Reference source not found.**], ellos declaran que un pronombre tiende a referirse al último sintagma nominal introducido. Afirmación que generaliza Allen [**Error! Reference source not found.**] para el resto de expresiones anafóricas denominándola recency constraint: tomar los antecedentes más cercanos que satisfacen todas las restricciones.

Competencia

A continuación se muestran en la Figura 8 los resultados asociados a la influencia de la competencia sobre el tipo de expresión anafórica. Esta competencia se mide en función del número de antecedentes posibles para una determinada expresión anafórica.

Competencia	***Expresión anafórica***			
	Descripción definida	***Descripción demostrativo***	***Pronombre personal***	***pronombre posesivo***
1	**5.9%**	**31.1%**	**88.4%**	**94.2%**
2	**10.5%**	**13.3%**	**6.4%**	**4.1%**
3	**83.6%**	**55.5%**	**5.2%**	**1.6%**
TOTAL	**100.0%**	**99.9%**	**100.0%**	**99.9%**
1= Existe un único antecedente posible **2= Existen dos posibles antecedentes** **3= Existen mas de dos antecedentes posibles**				

Figura 8. Relación entre competencia y expresión anafórica.

Esta tabla nos muestra que cuando existe un único antecedente, éste es altamente accesible para la expresión anafórica por lo que puede emplearse como medio de referencia anafórica en una expresión con poca carga léxica, como sería un pronombre. En el caso de que hubiese varios antecedentes la expresión anafórica debe tener rasgos más marcados para que la identificación del antecedente correcto sea posible.

Prominencia

La influencia de la prominencia del antecedente en el texto, se realiza dependiendo de dos factores: en primer lugar la función sintáctica que el antecedente tenga en la oración, y en segundo lugar la repetición o recurrencia que éste muestre a lo largo del texto. Las funciones sintácticas que marcan mayor prominencia en inglés Rico las ordena del siguiente modo: en primer lugar el

sujeto, a continuación el objeto directo y el objeto indirecto, y en último lugar los circunstanciales y los complementos regidos por el verbo. Mediante el segundo factor se permite la identificación de un antecedente que a pesar de no estar en posición sintáctica prominente consigue mediante la repetición hacerse prominente.

Unidad

A continuación en la Figura 9 se muestran resultados sobre la influencia de la unidad o localización de la expresión anafórica y su antecedente dentro de un mismo párrafo. Este concepto de unidad determina el efecto que la estructura discursiva tiene en la elección de las expresiones anafóricas. En el análisis el criterio de unidad se reveló como poco importante ya que la mayoría de las expresiones anafóricas se agrupan dentro de los valores 0 y 1, es decir, en el mismo párrafo o en el anterior.

Unidad	***Expresión anafórica***			
	Descripción definida	***Descripción demostrativo***	***Pronombre personal***	***pronombre posesivo***
0	**60.5%**	**73.3%**	**73.9%**	**91.0%**
1	**26.5%**	**22.2%**	**23.9%**	**8.1%**
2	**4.6%**	**4.4%**	**0.7%**	**0.0%**
3	**8.4%**	**0.0%**	**1.5%**	**0.8%**
TOTAL	**100.0%**	**99.9%**	**100.0%**	**99.9%**
0= El antecedente se encuentra en el mismo párrafo **1= El antecedente se encuentra en el párrafo anterior** **2= El antecedente se encuentra dos párrafos antes** **3= El antecedente se encuentra mas de dos párrafos antes**				

Figura 9. Relación entre unidad y expresión anafórica.

3.1.1.6 Información sobre la expresión anafórica

Como ya se vio anteriormente, las características del proceso de resolución de la anáfora varían en función del tipo de expresión anafórica. En la Figura 10 se

muestran los resultados del estudio realizado por Rico [**Error! Reference source not found.**] sobre la importancia de cada tipo de información en la resolución de diferentes tipos de anáfora en inglés. Como se puede observar, se confirman nuevamente las anteriores afirmaciones según las cuales los pronombres tienen menor carga léxica que las descripciones definidas, con lo que la información léxica y semántica es menos importante.

Tipo de información	***Expresión anafórica***			
	Descripción definida	***Descripción demostrativo***	***Pronombre personal***	***pronombre posesivo***
Sinonimia	**84%**	**82%**	**0%**	**8%**
Inferencia	**8%**	**9%**	**0%**	**0%**
Prominencia	**5%**	**0%**	**21%**	**10%**
Competencia	**0%**	**0%**	**15%**	**0%**
Distancia	**0%**	**0%**	**50%**	**81%**
Inf. Verbo	**3%**	**9%**	**13%**	**0%**
TOTAL	**100%**	**100%**	**99%**	**99%**

Figura 10. Información necesaria para resolver cada tipo de expresión anafórica.

3.1.1.7 Información obtenida a partir del estudio del corpus

La aplicación de información obtenida a partir de corpus al problema de la resolución de la anáfora está creciendo rápidamente. Por ello, cada vez es mayor la cantidad de corpus disponibles para la evaluación de los sistemas de tratamiento de la anáfora (por ejemplo los corpus: LOB, Susanne o el Penn Treebank). Estos corpus son de uso general, es decir, pueden estar anotados sintáctica y morfológicamente pero no están anotados respecto al discurso (roles temáticos o focos del discurso) ni respecto a las relaciones anafóricas entre diferentes constituyentes del corpus. La utilidad de los corpus anotados anafóricamente reside en poder utilizar información cuantitativa para restringir la

búsqueda de antecedentes en sistemas reales pobres en conocimiento[5] (knowledge poor systems).

McEnery, Tanaka y Botley [**Error! Reference source not found.**] nos muestran un poco acerca de los trabajos que actualmente se están realizando sobre la anotación de este tipo de información en corpus. Dentro de estos corpus y técnicas de anotación de los mismos destacan los siguientes:

- El UCREL Anaphoric Treebank éste consta de 100.000 palabras anotadas morfosintácticamente provenientes de artículos de diarios. Este corpus se desarrolló como parte de un proyecto de colaboración entre UCREL e IBM, y fue anotado utilizando la técnica de anotación de discurso IBM/UCREL desarrollada por Fligelstone [**Error! Reference source not found.**]. Aquí se recopilaron una amplia gama de características anafóricas y cohesivas. Estas características se asocian con unos determinados símbolos de anotación que codifican el tipo de relación y la dirección de la referencia en caso de ser relevante. Asimismo también es posible marcar grados de certeza en las anotaciones (tanto de la dirección de la referencia como del antecedente seleccionado), detección de múltiples antecedentes y rasgos semánticos para los pronombres en segunda y tercera persona.

  El inconveniente con este corpus es que sólo permite marcar relaciones anafóricas con antecedentes que aparecen explícitamente en el texto, y no cuando los antecedentes están relacionados con la expresión anafórica (por ejemplo cuando la expresión anafórica se refiere a un trozo extenso de texto).

- Rocha [**Error! Reference source not found.**] desarrolló una técnica de anotación, la cual se divide en tres pasos. Esta técnica es aplicada en los

[5] Tipo de sistema de tratamiento de la anáfora que restringe la cantidad de información necesaria para resolver la anáfora.

textos obtenidos de conversaciones en inglés y portugués, los cuales son anotados de acuerdo ala estructura de tópicos del texto (concepto similar al de la teoría del foco del discurso que se desarrollará más adelante).

En primer lugar Rocha establece para cada fragmento del discurso un tópico global o tópico del discurso, el cual puede permanecer durante todo el texto o cambiar en diferentes puntos del discurso. Este tópico estará formado por un sintagma nominal.

En segundo lugar divide el texto en segmentos de acuerdo a la continuidad del tópico local. Esto se realiza asignando un tópico de segmento local que es sólo válido en ese segmento, lo cual quiere decir que en el momento en que este tópico cambie también se considera que empieza un nuevo segmento. Estos cambios de segmento y de tópico local se anotan manualmente en el corpus.

Y en tercer lugar Rocha anota cada caso de anáfora que se produzca en el texto especificando cuatro propiedades: tipo de expresión anafórica, tipo de antecedente, grado de topicalidad del antecedente e información empleada para la resolución.

En este corpus se incluye algo nuevo que es intentar codificar la relación entre el tópico (o foco del discurso) y los casos de anáfora. De la misma manera clasifica las expresiones anafóricas y antecedentes de acuerdo a unos determinados criterios (las cuatro propiedades comentadas en el párrafo anterior), anotando además el tipo de conocimiento que considera necesario para resolver la anáfora. También este corpus está preparado para manejar otras relaciones al tener previsto la incorporación de textos multilingües. Sin embargo la principal desventaja de este corpus es que no utiliza estándares en los símbolos utilizados para la notación.

- Gaizauskas y Humphries [**Error! Reference source not found.**] superan el método propuesto por Rocha el cual es utilizar las etiquetas SGML

(*Standard Generalised Markup Language*) que para anotar los casos de anáfora. Dichas etiquetas tienen la ventaja de ser actualmente un estándar para la codificación electrónica de textos que permite su intercambio entre diferentes sistemas. Pero tienen la desventaja de sólo permitir un número determinado de expresiones anafóricas y también presenta el inconveniente de su diseño orientado a tareas de resolución automática, donde el éxito de cada anotación ha de ser medido, es decir, no se diseñó con el objetivo de etiquetar grandes corpus.

- Botley [**Error! Reference source not found.**] desarrolla una técnica la cual nos sirve para describir las distintas formas en que las expresiones para demostrativas funcionan anafóricamente en textos escritos y hablados. Para esto el clasifica las expresiones anafóricas demostrativas de acuerdo a cinco criterios:

    1. El grado de recuperabilidad del antecedente
    2. Dirección de la referencia
    3. Tipo de referencia (endófora o exófora)
    4. Función sintáctica
    5. Tipo de antecedente

  Este método al igual que el de Rocha presenta la ventaja de permitir marcar mayor información sobre los casos de anáfora que el método de UCREL.

En el mismo trabajo de McEnery, Tanaka y Botley [**Error! Reference source not found.**] se menciona cómo se está trabajando en la universidad de Lancaster en colaboración con IBM Yorktown Heights en el desarrollo de un corpus en el cual van a estar anotadas las referencia anafóricas. En este momento el corpus aún no está terminado, si embargo ya muestran algunos de los resultados obtenidos en el

análisis. En los cuales destacan que efectivamente los antecedentes de las expresiones anafóricas estudiadas presentan un modelo común, el cual varía en función del tipo de texto que se trate. Esta información puede resultar muy útil para adaptar la estrategia general de resolución anafórica al tipo de texto que se esté tratando, aunque tal y como expresan McEnery, Tanaka y Botley todavía estamos lejos de tener a nuestra disposición tal información.

La información que se obtuvo del análisis del corpus ya se aplica en otros trabajos, especialmente para obtener reglas de resolución especializadas para cada tipo de expresión anafórica y personalizadas para cada tipo de texto. Para obtener dichas reglas es utilizada la información probabilística obtenida tras el estudio del corpus.

El principal problema es que si el corpus no esta anotado adecuadamente existe dificultad para obtener la información anafórica de éste. Además, en la actualidad esta anotación se está llevando a cabo prácticamente de manera manual, por lo que hace más lento su progreso y su posterior empleo en aplicaciones reales de procesamiento del lenguaje natural. Mitkov [**Error! Reference source not found.**] propone un trabajo con el que pretende superar esta dificultad, en el que esboza un sistema para la anotación semiautomática de la anáfora pronominal en corpus. Para ello utiliza un sistema de resolución pobre en conocimiento, seguido de una posterior fase de anotación y revisión manual de los resultados obtenidos. Mitkov [**Error! Reference source not found.**] detalla este sistema de resolución en su mismo trabajo, el cual se desarrollará mas detalladamente en la siguiente subsección, la cual trata de los métodos alternativos de resolución de anáfora. Brevemente se puede decir que este sistema trabaja sobre la salida de un etiquetador[6], sin realizar un análisis sintáctico, y utiliza una serie de indicadores de antecedente, es decir, un grupo de reglas heurísticas dependientes del dominio que sirven para la detección del antecedente correcto. El dominio sobre el que trabaja es el de manuales de informática y equipos de sonido de alta fidelidad.

[6] Part of speech tagger o también conocido como POS tagger.

Este sistema de anotación semiautomática no está desarrollado actualmente y se plantea como una posible futura colaboración entre la Universidad de Lancaster, British Telecom y la Universidad de Wolverhampton.

3.1.2 Algoritmos

Se vieron detalladamente y por separado cada fuente de información utilizada en el tratamiento de la anáfora, por lo tanto ahora se describirá los distintos algoritmos o sistemas que hacen uso de estas fuentes. Estos algoritmos difieren principalmente en la cantidad o tipo de información que incluyen, o bien en el modo que organizan o coordinan estos tipos de información. Por esto, se comenzará por describir las primeras estrategias de resolución de la anáfora que comparten la característica común de limitarse prácticamente a una sola fuente de información. En seguida se dividirán los sistemas restantes en dos grupos en función del tipo de información que emplean: sistemas integrados y sistemas alternativos. La diferencia fundamental entre ambos es que los alternativos se basan en la información obtenida a partir del estudio de corpus, mientras que en los integrados utilizan el resto de fuentes de información. A su vez el grupo de sistemas integrados lo dividiremos en cuatro subgrupos en función del modo en que coordinan las diferentes fuentes de información:

- Sistemas democráticos basados en restricciones y preferencias.
- Sistemas democráticos basados únicamente en preferencias.
- Sistemas pobres en conocimiento (knowledge poor systems).
- Sistemas consultivos basados en la teoría del foco del discurso.

Los primeros tres subgrupos dan la misma importancia a cada fuente de información, mientras que en el último surge una fuente de información que propone candidatos a antecedente, y las restantes fuentes de información se limitan a confirmar o rechazar estos candidatos.

3.1.2.1 Primeras estrategias al tratamiento de la anáfora

A continuación se verán en resumen las primeras estrategias al problema de la anáfora. También se mostrarán a la primera generación de investigaciones en el tratamiento del lenguaje natural y la resolución de la anáfora, cuya característica común es la de centrarse prácticamente en una sola fuente de información, y basándose principalmente en la utilización de reglas heurísticas para la resolución de ciertos casos de anáfora.

Alrededor de los años sesenta y setenta se realizaron las primeras estrategias al problema de la anáfora, estas tienen un principal inconveniente el cual es que tienen a centrar la búsqueda de posibles antecedentes a aquellos encontrados en la misma oración. Para lo cual hacen uso únicamente de conocimiento sintáctico para localizar el antecedente, por lo que se encuentran un gran número de casos de anáfora que no pueden tratar.

El sistema STUDENT es un ejemplo de estas primeras estrategias desarrollado por Bobrow [**Error! Reference source not found.**]. Este sistema resuelve problemas de álgebra e incorpora unos pocos procedimientos heurísticos para resolver algunos tipos de anáfora y repeticiones incompletas. Por ejemplo, es capaz de detectar el antecedente del pronombre en el siguiente texto: *The number of soldiers the Russians have is half the number of guns they have. The number of guns is 7000. What is the number of soldiers they have?*. Sin embargo, estos procedimientos son fácilmente burlados ya que la oración no se analiza sintácticamente.

Weizenbaum [**Error! Reference source not found.**] desarrolló un programa llamado ELIZA con el que ocurre algo parecido. Éste es uno de los primeros programas escritos para simular el comportamiento humano, el cual es capaz de imitar a un psiquiatra que dialoga con su paciente. Este programa no es un sistema experto con conocimiento de psiquiatría, sino que sencillamente se trata de un emparejado de estructuras (*pattern-matching*) con el objetivo de enlazar con

la conversación del paciente (que el programa sea capaz de continuar la conversación a partir de la última frase del paciente). En cuanto al tratamiento de la anáfora se reduce al intercambio entre pronombres (por ejemplo será capaz de cambiar de primera y segunda persona, es decir, cuando el usuario utiliza el pronombre yo, al contestar ELIZA lo intercambia por el pronombre tú).

Winograd [**Error! Reference source not found.** y **Error! Reference source not found.**] propone el sistema interactivo SHRDLU que permite al usuario entablar un diálogo con el ordenador para obtener información acerca de determinadas cuestiones. Los diálogos que el usuario puede mantener con el programa giran en torno a un campo temático concreto: el mundo de las figuras geométricas y las posiciones, formas y colores que éstas pueden tomar. Además SHRDLU tiene la posibilidad de aceptar instrucciones y ejecutarlas, para ello dispone de un brazo de robot para mover los bloques, una mesa donde se colocarán estos bloques y los bloques propiamente dichos. Este sistema puede tratar algunos pronombres personales, sintagmas nominales definidos y determinados casos de anáfora tipo "one".

Se compone de tres módulos: un analizador sintáctico o parser, un analizador semántico y un módulo que incorpora técnicas de razonamiento que permite resolver las cuestiones más complejas. Éste incorpora técnicas heurísticas mucho más complejas que las desarrolladas en el sistema STUDENT, permitiendo referencias a partes anteriores de la conversación entre el programa y el usuario. Para ello, almacena en una pila todos los grupos nominales que encuentra asignándoles a cada uno un valor numérico que indique la preferencia de la expresión anafórica por éste, y en caso de empate se solicita la intervención del usuario para determinar la referencia. Winograd utiliza la concordancia en número, género y persona entre la expresión anafórica y su antecedente, eligiendo el antecedente más reciente que pase estas restricciones de concordancia. Algunas de las reglas que emplea son fácilmente de engañarles, como sería la siguiente: "En el caso que aparezcan dos veces los pronombres *it* o *they* en la misma oración o en dos oraciones adyacentes, se considerará que estos pronombres son

correferenciales", la cual no se cumple en la siguiente frase: *He put the box on the table. Because it$_i$ wasn't level, it$_j$ slid off*.

Como aspectos positivos de este sistema, se encuentra que es capaz de detectar e interpretar correctamente expresiones anafóricas que forman parte de su propio antecedente, como por ejemplo en: *... a block which is bigger than anything which supports it*. También maneja ciertos usos de one con significados opuestos, como los casos de pares de palabras como grande y pequeño, que se usan frecuentemente como opuestos, por ejemplo en *a big green pyramid and a little one*, aquí la expresión *little one* significa *little green pyramid* y no *little pyramid* o *little big green pyramid*. Sin embargo esta detección fallaría en el caso de adjetivos que no tengan significados opuestos, como en el caso de: *a big blue one*, *a big green one and a little blue one*.

Woods [**Error! Reference source not found.** y **Error! Reference source not found.**] describe un interfaz en lenguaje natural para una base de datos sobre minerales procedentes de las piedras traídas de la luna por la expedición Apolo-11. Este sistema se le conoció como LSNLIS (Lunar Sciences Natural Language Information System) o también como LUNAR. Tiene tres componentes básicos que procesan el lenguaje natural: un analizador sintáctico, un módulo de interpretación semántica y un último módulo encargado de la gestión de la base de datos. De modo similar al SHRDLU, para realizar el tratamiento de la anáfora almacena cada entidad que aparece en el discurso junto con sus respectivas representaciones sintácticas y semánticas. Distingue dos tipos de anáfora: parcial y completa. La anáfora parcial se caracteriza por el empleo de un pronombre que se refiere a una parte de un sintagma nominal aparecido con anterioridad, por ejemplo en: *Give me [all analyses of sample 10046 for hydrogen]$_1$. Give me [them for oxygen]$_1$*. Estos casos los detecta por la presencia de una oración de relativo o un sintagma preposicional que modifica al pronombre (for oxygen). Para resolverlos efectúa una búsqueda entre todos los grupos nominales que han sido mencionados anteriormente, e intenta localizar aquél que tenga una estructura sintáctica y semántica paralela a la expresión anafórica empleada. En la anáfora

completa aparece un pronombre que se refiere a un sintagma nominal completo, tal y como ocurriría en: *Which [coarse-grained rocks]$_i$ have been analyzed for cobalt? Which ones$_i$ have been analyzed for strontium?*. La estrategia de resolución depende de si la anáfora se construye mediante un demostrativo o un pronombre. Cuando la expresión anafórica la constituye un sintagma nominal introducido por un demostrativo, el método aplicado consiste en buscar un sintagma nominal cuyo nombre sea el mismo que acompaña al demostrativo: *Do any breccias$_i$ contain aluminium? What are [those breccias]$_i$?*. En el caso que la anáfora se construya mediante un pronombre, entonces obtendrá información semántica de la frase en que aparece, información que procesará su módulo de interpretación semántica.

Este sistema tiene las mismas limitaciones que los anteriormente descritos, es decir, sólo puede resolver expresiones anafóricas que tienen la misma estructura que sus antecedentes, por ejemplo en *Give me [all analyses of sample 10046 for hydrogen]$_1$* ninguna de las dos siguientes frases se podría detectar como anáfora parcial: *Give me [the oxygen ones]$_1$*. *Give me [those that have been done for oxygen]$_1$*. Además de ello, también tiene la limitación de no poder tratar con la anáfora intrasentencial ya que los sintagmas nominales de esa oración no estarán disponibles hasta que se haya analizado completamente la oración.

Charniak [**Error! Reference source not found.**] detalla un sistema cuyo dominio son historias infantiles. Éste tiene como objetivo principal estudiar el tipo de inferencias sobre el mundo real que son necesarias para hacer comprender al ordenador pequeñas historias, contestar a preguntas sobre estas historias y resolver problemas de ambigüedad anafórica. Trabaja sobre sintagmas nominales definidos y pronombres. Para realizar inferencias y razonamientos el sistema emplea un método esencialmente heurístico. La estrategia se basa en la aplicación de los llamados *demons* o *rutinas* que codifican hechos sobre el mundo que deben satisfacerse para llegar a la comprensión del texto. Estas rutinas se van activando a medida que se procesa la historia y aparecen conceptos nuevos. Para resolver anáforas, en primer lugar construye una lista de posibles

antecedentes de la expresión anafórica, a continuación, si se comprueba que esta lista contiene un único candidato, éste se acepta como antecedente; en el caso de que la lista contenga más de uno, entonces la expresión anafórica se representa como una variable y se aplican diferentes demons con esta variable a la lista de antecedentes; finalmente, la variable que representa a la expresión anafórica quedará ligada a aquel antecedente que sea válido para el demon que contiene la variable. Por ejemplo, para el siguiente texto: *Janet put some money on the sink. Mother said, 'If you leave the money there it may fall in the drain'*, para encontrar el antecedente del pronombre *it*, el sistema consulta la lista de posibles antecedentes, en este caso *money* y *sink*. Para decidirse por uno de ellos, convierte al pronombre en una variable y la incluye dentro del demon que represente la información *it may fall in the drain*, interpretándose para ambos candidatos. Obviamente se obtiene como solución que el dinero sí que puede colarse por el desagüe, mientras que no ocurre así con el propio fregadero.

La desventaja con la que cuenta este sistema es que a la hora de aplicar varios demons para resolver una relación anafórica siempre tiene preferencia aquel que se haya empleado antes, lo cual no produce habitualmente resultados correctos. Además tiene el inconveniente de que aunque el dominio sea restringido se necesita gran cantidad de demons para codificar esta información.

Wilks [**Error! Reference source not found.**] elabora un sistema traductor de inglés a francés en el cual implementa su teoría de la semántica de preferencias. En la cual utiliza patrones semánticos para interpretar las palabras en sus contextos. Por lo que concibe el texto como un conjunto de bloques semánticos. De esta manera, el se convierte en una unidad semántica que no necesita la sintaxis para su análisis. Cada bloque semántico o texto está compuesto por plantillas, las cuales se unen mediante patrones y reglas de inferencia de sentido común. El vínculo de unión entre estos elementos lo constituyen las fórmulas. Este sistema hace uso de cuatro niveles de resolución de la anáfora pronominal dependiendo del tipo de pronombre y del mecanismo para resolverlo. En el primer nivel, se aplican las plantillas y patrones existentes, utilizando únicamente

conocimiento de las palabras que aparecen en el texto. Por ejemplo, en: *Give the bananas$_i$ to the monkeys$_j$ although they$_i$ are not ripe, because they$_j$ are very hungry*, cada pronombre *they* es interpretado correctamente utilizando el conocimiento de que los monos son seres animados, por lo que tienen la posibilidad de estar hambrientos, y los plátanos es una fruta la cual puede estar o no madura. En el segundo nivel, se construyen nuevas plantillas semánticas que de alguna manera están implícitas en las plantillas que ya existen. Este nivel se aplicará en caso que el nivel anterior falle en encontrar un único antecedente del pronombre, por lo que se hace necesario utilizar métodos de inferencia para obtener conocimiento del mundo real. El tercer nivel se aplicará en caso que todavía haya más de un antecedente, en cuyo caso se intenta encontrar el tópico o foco de la oración que se considerará como antecedente. Este nivel supone el empleo de reglas de inferencia sobre el mundo real que van más allá de las definiciones y significados codificados en las fórmulas. En caso que los tres niveles anteriores fallen entonces es necesario seleccionar el antecedente por defecto (nivel cuatro), lo cual, equivale a suponer que todavía se está hablando de lo mismo que se había hablado hasta ese momento.

En los setentas se realizaron algunos estudios sobre el uso de conocimiento de dominio en el discurso. Los sistemas que se hicieron en ese tiempo no son de uso general. Algunos de éstos sistemas son GUS [12], SAM [27] y el Task Dialogue Understanding System de Grosz [**Error! Reference source not found.**]. Donde este último lleva acabo la distinción entre conocimiento del dominio, información del discurso y la intención del mismo. Grosz construye este sistema como parte su tesis doctoral. En la cual su trabajo trata con la estructura del discurso y la noción el foco del discurso más que con el tema de la resolución de la anáfora, aunque supone que una correcta formulación de la estructura del discurso contribuye a la comprensión de la anáfora. Esta formulación de la estructura del discurso se englobará dentro de la denominada teoría del foco del discurso, que se desarrollará con mayor profundidad posteriormente.

Hobbs [**Error! Reference source not found.**] desarrolló dos estrategias para resolver las referencias pronominales. En la primera hace uso únicamente de la información sintáctica, mientras que en la otra realiza un análisis semántico del texto. En la primera estrategia Hobbs desarrolla un sencillo algoritmo que trabaja sobre los árboles sintácticos de las oraciones del texto. Estos árboles sintácticos representan la estructura gramatical del texto y se utilizan para realizar la búsqueda de los sintagmas nominales, antecedentes de una determinada expresión anafórica, para ello utilizará las restricciones típicas como la concordancia en número, género y persona, este algoritmo lo comprobó eligiendo 100 frases con distintas ocurrencias de los pronombres *he*, *she*, *it* y *they*, logrando un porcentaje de éxito del 81.1%. A pesar de este porcentaje de éxito Hobbs considera que es necesario añadir la información semántica al algoritmo, ante los casos en los fracasa y de esta manera conseguir el 100% de éxito,

En la segunda estrategia, Hobbs hace la descripción de un sistema que comprende algunas operaciones semánticas para desarrollar ciertos mecanismos de inferencia a partir de una base de conocimiento en la que se almacenan datos del dominio del texto. Aparecen cuatro operaciones básicas a detectar o verificar relaciones entre oraciones, interpretación de predicados, eliminación de redundancias e identificación de entidades. Estas operaciones se diseñan con el objetivo de reconocer la estructura e interrelaciones implícitas en el texto. Esta estrategia presenta problemas computacionales, como por ejemplo la complejidad exponencial en los procedimientos de búsqueda. Otro inconveniente de este algoritmo es precisamente la ausencia del módulo de análisis sintáctico que construya el árbol sintáctico correcto para todas las oraciones.

Webber [**Error! Reference source not found.**] en su tesis doctoral describe una estrategia computacional de la resolución de la anáfora en la que considera el discurso como una colección de diferentes tipos de entidades (individuales, conjuntos, acontecimientos o acciones). Estas entidades del discurso serán los antecedentes de las expresiones anafóricas. En el modelo Webber aparecen lo que él denomina invoking descriptions (*ID*) de cada entidad del discurso, que

significan la primera mención de una entidad del discurso, y las expresiones anafóricas tendrán como antecedentes estas ID. Webber trata con la anáfora de tipo “one”, pronombres y algunos casos de cuantificación.

Alshawi [**Error! Reference source not found.**] desarrolló un sistema para la resolución de la anáfora, este se basa en dos mecanismos esenciales: el de memoria y el contextual. Dichos mecanismos se emplean para la identificación de antecedentes, resolución de ambigüedades, y la generación de enunciados a partir de un texto y almacenarlos en una base de datos.

El primer mecanismo, que es el de memoria, es utilizado para almacenar información a medida que el sistema va analizando el texto. Para poder realizar su trabajo, este emplea una serie de reglas semánticas formales las cuales se definen mediante dos tipos de función (*ref* y *rel*) y dos tipos de afirmación (specialisation y corresponds). La función “ref” establece una referencia entre una entidad y el conjunto de objetos del mundo que esta representa. Por ejemplo, la regla ref (*COMPUTER*) crea la referencia entre la entidad computer y los objetos del mundo real que corresponden a ella. La función rel establece una relación entre un par de entidades y el conjunto de pares de objetos que corresponden a éste en el mundo real. Respecto a los tipos de afirmación, la specialisation indica la referencia de un subconjunto a un conjunto mayor y la corresponds expone la relación que existe entre un par de entidades y otro par de entidades correspondientes a un conjunto mayor.

El segundo mecanismo, es el contextual, el cual nos señala la información más relevante en el texto. Para esto se definen siete factores contextuales que son aplicados al texto conforme se va analizando.

- Factor que es encargado de señalar la oración más reciente en el texto.
- Factor que indica qué párrafo es el más reciente.

- Factor que indica la relevancia del sujeto en las oraciones construidas con el verbo *to be* y las oraciones de pasiva (énfasis sintáctico).

- Factor que indica el camino o historia que han seguido las operaciones del mecanismo de memoria.

- Factor que señala el uso de deíxis[7] (que sirve para dirigir la atención del receptor desde un interés pre-existente a otro nuevo).

- Factor de asociación entre nodos semánticos cuyos vecinos hayan sido activados anteriormente.

- Factor que asocia un enunciado creado en la base de datos con todas las entidades correspondientes.

Cada factor consiste en un grupo de entidades a las que se asocia un valor numérico que determina su peso o importancia en el conjunto del texto. Para localizar un antecedente en una relación anafórica se lleva a cabo una búsqueda entre todas las entidades cuyo peso supere un umbral determinado. Si sólo existe una entidad que supere este umbral se aceptará esta como antecedente. Si son varias las entidades propuestas se tomará como antecedente aquella cuyo peso sea mayor.

3.1.2.2 Sistemas integrados basados en el conocimiento

En la anterior subsección se revisó en breve las primeras estrategias al problema de la anáfora en las que no se trata de manera conjunta cada una de las fuentes de información vistas anteriormente, basándose principalmente en la utilización de reglas heurísticas para la resolución de ciertos casos de anáfora. A continuación

---

[7] El termino deixis es utilizado para designar la codificación en los enunciados de los datos referidos al contexto situacional, es decir, del lugar, tiempo y participantes. Tienen valor deíctico los pronombres demostrativos y personales, los adverbios de lugar, tiempo y modo y los tiempos verbales.

estudiaremos los sistemas que utilizan diversas fuentes de conocimiento con el objetivo de tratar la anáfora. A estos sistemas se los puede encontrar en la literatura como sistemas integrados basados en el conocimiento (*integrated knowledge-based systems* o sencillamente *integrated systems*). Estos sistemas se diferencian de los sistemas alternativos, que expondremos en la siguiente subsección, porque no hacen uso de información obtenida a partir de corpus.

Los sistemas integrados pueden ser clasificados según el modo en que utilizan o coordinan estas fuentes de información. Según Carter [**Error! Reference source not found.**] y Rico [**Error! Reference source not found.**] se distinguen dos enfoques para la coordinación de todas estas fuentes: democrático y consultivo.

Enfoque democrático, en este todas las fuentes de conocimiento tienen asignado el mismo papel en la selección del antecedente anafórico, es decir, que aquellas entidades que pueden ser susceptibles de convertirse en antecedente pueden surgir por igual de aportaciones de la información morfológica, sintáctica, semántica o pragmática. Un claro ejemplo de enfoque democrático es el propio trabajo de Rico [**Error! Reference source not found.**] justificando este enfoque según sus propias palabras, “porque de este modo los valores que aporta cada fuente de información son siempre relativos ya que no podemos decidir de antemano qué fuente es la que aportará la información decisiva”. Dentro del enfoque democrático distinguiremos cuatro apartados. En los primeros dos se estudiaran diversos ejemplos de sistemas democráticos agrupados en función de cómo se considere cada fuente de información: como restricciones y preferencias, o sólo como preferencias. En el siguiente apartado se agrupan aquellos sistemas democráticos aplicados a la resolución de la anáfora pronominal que comparten la característica común de restringir la cantidad de información que necesitan para el tratamiento de este tipo de anáfora.

En el enfoque consultivo, solamente una fuente de información está habilitada para proponer antecedentes, y el resto de fuentes de información tienen la función de confirmar o rechazar los antecedentes propuestos. Un ejemplo habitual de este

segundo enfoque es el de los algoritmos basados en la teoría del foco del discurso, los cuales veremos en profundidad en el último apartado.

SISTEMAS DEMOCRÁTICOS BASADOS EN RESTRICCIONES Y PREFERENCIAS

Carbonell y Brown [**Error! Reference source not found.**] ellos se muestran de acuerdo con la necesidad tratar la anáfora a profundidad como un fenómeno en el que intervienen muchas y diferentes fuentes de información. Dividen los tipos de conocimiento en dos: restricciones y preferencias.

Las restricciones se utilizan para eliminar candidatos, para las cuales Carbonell y Brown elegirían las siguientes fuentes de información: morfológicas, semánticas (eliminan todas aquellas entidades discursivas cuyos rasgos semánticos no concuerden con las restricciones semánticas exigidas por la expresión anafórica) y restricciones impuestas por la acción que se realiza en la oración. La última restricción va más allá de la anterior semántica, ya que utiliza información inherente de la acción que realiza el verbo, por ejemplo en *Juan le dio a Pedro una manzana. Él se la comió*, el verbo dar impone la condición de que la persona que da algo a alguien deja de poseer ese algo. Si el antecedente de *él* fuera *Juan* se crearía un conflicto entre la condición del verbo *dar* y la de *comer*, que establece que para que alguien coma algo es necesario que ese alguien posea ese algo.

Las preferencias ayudan a facilitan la selección del antecedente correcto indicando qué candidatos se consideran mejores que otros. Como criterios de preferencia utilizaron la preferencia de la expresión anafórica por aquellos candidatos que se encuentren en posiciones paralelas, la preferencia por los antecedentes que aparecen en posiciones temáticas y preferencia por los antecedentes que se encuentren en la misma oración que la expresión anafórica.

Carbonell y Brown consideran que la preferencia por candidatos que se encuentren en posiciones paralelas se refiere tanto a paralelismo sintáctico, semántico o pragmático. Como posiciones sintácticas paralelas entienden aquellas

en las que la expresión anafórica y su antecedente desempeñan la misma función sintáctica aunque la unidad lingüística que las representa sea distinta: *The robot gave the dog$_i$ a bone. John also gave it$_i$ some water*. Por posiciones semánticas paralelas consideran las relaciones que se dan entre una expresión anafórica y un antecedente que comparten la misma categoría semántica, aunque se manifiesten en diferentes posiciones sintácticas en la oración: *Juan llevó la caja de Pedro a Luis$_i$. Él también le$_i$ llevó los libros de María*. Aquí *le* y *Luis* comparten el rasgo semántico de receptor. Y finalmente por posición pragmática interpretan aquella que obliga a seleccionar como antecedente el que ayude a mantener la cohesión discursiva: *María condujo del parque al club. Pedro fue allí también*, la primera parte del discurso crea una acción de movimiento del parque al club. En la segunda oración debemos interpretar *allí* como refiriéndose al *club* si queremos mantener la cohesión discursiva.

En cuanto a la preferencia por los antecedentes que aparecen en posiciones temáticas se establece que debe darse mayor prioridad a los posibles antecedentes que representen los tópicos o temas del discurso. Por ejemplo, en: *María dijo a Ana de ir a Nueva York. Por qué hizo ella esto?.* Aquí *María* está en posición temática por lo que se le prefiere como antecedente de *ella*.

El sistema propuesto por Carbonell y Brown tiene un módulo de tratamiento de las relaciones anafóricas que actúa después del análisis sintáctico y semántico con la siguiente secuencia de pasos:

- Considera como posibles antecedentes todos los sintagmas nominales que hayan aparecido con anterioridad.
- Aplica sobre ellos las restricciones y elimina así todos los candidatos que no las cumplan.
- Sobre los candidatos que quedan aplica las preferencias.

- Como resultado, cada candidato tendrá asignado un peso o valor correspondiente a la relevancia que le otorga cada preferencia. Por último, el sistema selecciona como antecedente el que mayor peso haya obtenido.

En [**Error! Reference source not found.**] Rich y LuperFoy siguen la misma línea de trabajo de Carbonell y Brown, afirmando que deben crearse diferentes teorías para explicar los diversos problemas y aspectos que componen el tratamiento de la anáfora. Para ello, crean distintos módulos que se corresponden a su vez con diferentes teorías de la anáfora. Cada uno de estos módulos aborda un aspecto distinto de la identificación de antecedentes, e impone una serie de restricciones, por lo que reciben el nombre de fuentes de restricción (*constraint sources o CS*). Aparte de cada uno de estos módulos independientes hay un módulo superior, el controlador, que coordina las propuestas e intercambios entre todos los módulos. Cada CS tiene cuatro tipos de funciones:

- Modelado: es la función que mantiene el modelo de interpretación creado por cada CS. Por ejemplo, supongamos que el módulo activo es el que contiene información sobre el discurso. Para poder aplicar las restricciones sobre los posibles antecedentes se construye un modelo de interpretación discursiva que permanece activo mientras es necesario.

- Exhibición de restricciones: establece todas las restricciones que interactúan en la identificación de antecedentes asegurando así que la interpretación será correcta.

- Formulación de hipótesis: se encarga de proponer una lista de candidatos a antecedente mediante la asociación de una puntuación a cada uno de ellos en función de los resultados obtenidos tras aplicar las restricciones.

- Evaluación: evalúa los antecedentes propuestos por la función anterior.

Esta estrategia se lleva a la práctica en LUCY, un sistema de comprensión de textos. Para asegurar la independencia entre los diferentes módulos utilizan la

llamada arquitectura de pizarra. Para ilustrar el funcionamiento de esta arquitectura imaginemos un grupo de especialistas dedicados a resolver un problema determinado. Todos ellos están reunidos ante una pizarra con el objetivo de comunicar sus conocimientos y buscar una solución al problema. Por turnos cada uno escribe en la pizarra su contribución, de modo que el siguiente especialista pueda utilizarla y añadir nuevos datos. Este proceso continúa hasta que el problema está resuelto o ninguno de los especialistas puede añadir más información. Esta arquitectura se ha utilizado sobre todo en reconocimiento del habla, y se caracteriza porque utiliza diversas fuentes de información para proponer y evaluar diferentes hipótesis.

El algoritmo utilizado se resume en estos pasos:

1. Análisis sintáctico y semántico.
2. Actualización de la estructura discursiva.
3. Se activan las funciones del modelado y el de exhibición de restricciones.
4. Se activa el controlador de la anáfora, que a su vez, llama al de formulación de hipótesis y al de evaluación.
5. El resultado es una lista de posibles antecedentes ordenados según la puntuación obtenida tras la aplicación de los CS.

Para la asignación de puntuaciones a cada candidato, emplean una función que calcula la media entre la puntuación absoluta que otorga cada CS a cada candidato (esta puntuación se sitúa entre -5 y +5) y el índice de confianza o seguridad de que el antecedente elegido es correcto (entre 0 y 1).

SISTEMAS DEMOCRÁTICOS BASADOS ÚNICAMENTE EN PREFERENCIAS

En el anterior apartado hemos visto sistemas que clasifican las fuentes de información que intervienen en el tratamiento de la anáfora en restricciones y preferencias. Las restricciones tienen como objetivo eliminar antecedentes para una determinada expresión anafórica, mientras que las preferencias ordenan los antecedentes que han quedado tras aplicar las anteriores restricciones para seleccionar el antecedente preferido. En este apartado veremos otra alternativa de manejo de estas fuentes de información en la que se eliminan las restricciones, considerándose todos los tipos de información como preferencias.

Un ejemplo de sistema basado únicamente en preferencias es el propuesto por Rico en su tesis doctoral [**Error! Reference source not found.**]. Este sistema está basado en el producto escalar de vectores y se fundamenta en los siguientes tres procesos:

- Tratamiento simultáneo de toda la información lingüística (sintáctica, morfológica, semántica y pragmática).
- Asignación de relevancia a cada fuente de información en función del contexto lingüístico.
- Comparación en términos de igualdad de cada antecedente posible y su expresión anafórica, es decir, todas las entidades discursivas tienen las mismas posibilidades de ser consideradas como antecedente, por lo que no se excluye a priori ningún candidato (todas las fuentes de información se consideran como preferencias).

Esta estrategia parte de la idea de codificar de forma numérica cada fuente de información representándola en forma de vector. De este modo cada expresión anafórica y posible antecedente tendrán asignado un vector numérico que simboliza los valores que tienen para cada fuente de información. Para realizar la comparación entre dos de estos vectores utilizará el producto escalar, ya que el número que se obtiene de este producto permite fijar la posición relativa entre ambos, permitiéndonos un mecanismo de comparación entre vectores. Para saber

qué posición relativa tiene un vector respecto del otro necesitamos conocer el ángulo que separa ambos para lo cual se utiliza la fórmula mostrada en la Figura 11.

$$v \cdot w = \sum_{i=1}^{n} v_i \cdot w_i$$

$$\cos\theta = \frac{v \cdot w}{\|v\| \cdot \|w\|} \quad , \quad \|v\| = \sqrt{\sum_{i=1}^{n} v_i^2}$$

Figura 11. Cálculo del ángulo de separación entre dos vectores basado en el producto escalar entre ambos.

La codificación de las distintas fuentes de información mediante vectores ya se utilizó por Sutcliffe [**Error! Reference source not found.**] para representar significados de palabras con el objetivo de llevar a cabo comparaciones entre distintas palabras. Después de realizar el cálculo del ángulo de separación entre ambos vectores, el que tenga menor ángulo indica que sus significados comparten muchos rasgos. Para realizar esta codificación Rico define un conjunto de atributos anafóricos aplicables de manera general tanto a entidades discursivas como a expresiones anafóricas, los cuales deben cumplir las siguientes condiciones:

1. Deben contener toda la información lingüística necesaria para consolidar una relación anafórica.

2. Se deben poder codificar en forma de valores numéricos.

3. El conjunto de atributos debe ser el mismo tanto para las expresiones anafóricas como para las entidades discursivas.

4. Debemos asignar los valores siguiendo siempre los mismos parámetros y criterios independientemente de que estemos tratando con una expresión anafórica o con una entidad discursiva.

5. El conjunto de atributos debe ser finito y abarcar todos los rasgos posibles.

6. El conjunto de atributos debe representar con exactitud el tipo de información necesaria para la identificación de antecedentes, y sólo dicha clase de información.

Rico optará por los siguientes atributos anafóricos:

- Información morfológica: género (neutro, masculino, femenino), número (singular, plural), persona (primera, segunda, tercera).

  Aquí resuelve el problema de una referencia a un nombre colectivo, según el cual la expresión anafórica con la que establecen referencia puede estar tanto en singular como en plural. Para resolver este problema el atributo de número del nombre colectivo lo pone tanto en singular como en plural, dejando que sean otras fuentes de información las que actúen de manera decisiva.

- Información sintáctica: sujeto, objeto (directo, indirecto), complemento (circunstancial, preposicional).

- Información semántica:

  1. Rasgos semánticos de carácter general que aseguran la consistencia entre antecedente y expresión anafórica: humano / no humano, animado / no animado, abstracto / no abstracto, objeto / no objeto, contable / incontable, sinonimia e hiperonimia.

  2. Rasgos que aseguran las restricciones de selección (dependen del tipo de verbo). Por ejemplo, el verbo de la expresión anafórica *modernize them* impone la restricción de que su objeto directo sea un objeto físico, por ello en el

vector de *them* aparecerá el rasgo objeto físico en este campo.

- Información pragmática:

    1. Prominencia (alta, baja)

    2. Distancia oracional (misma oración, diferente oración), distancia clausal (misma cláusula, diferente cláusula).

Ejemplo de distancia. Para la expresión anafórica *himself* se exige que esté en la misma cláusula y en la misma oración, por ello su vector contendrá los siguientes valores: (distancia oracional = misma oración, distancia clausal = misma cláusula). Para los posibles antecedentes, se les pondrá uno u otro valor en función de si están en la misma oración o cláusula que la expresión anafórica. Sin embargo, para la expresión anafórica *him* se exige que esté en distinta cláusula y misma oración por lo que su vector será: (distancia oracional = misma oración, distancia clausal = diferente cláusula). Ejemplo de prominencia. *Sandy walked her dog near a bull one day. It walked quietly along*, aquí dog tendrá más prominencia que *bull*, por lo que tendrá valor alto, mientras que *bull* tendrá valor bajo, y el pronombre *it* tendrá valor alto.

Respecto a los grados de aplicación de cada atributo se calcularán en función de la información lingüística que se extrae de la expresión anafórica y su antecedente. Para ello, se codificarán los grados de aplicación mediante valores numéricos, de tal modo que si un atributo se aplica en mayor grado que otro el valor correspondiente será más alto. Además clasificará los atributos anafóricos en dos tipos: restrictivos y no restrictivos. Para los primeros, su aplicación restringe las características que la entidad discursiva y la expresión anafórica pueden tener. Tendríamos como ejemplos los atributos morfológicos, semánticos y el atributo pragmático de distancia clausal. Los segundos, los atributos no restrictivos, no imponen ningún tipo de restricción a la relación que la entidad discursiva y la

expresión anafórica puedan establecer. Como ejemplos aparecen los atributos sintácticos, la prominencia y la distancia oracional.

En cuanto a la asignación de valores numéricos a cada atributo, se definen dos reglas: una para los atributos restrictivos y otra para los no restrictivos. Para la primera, cualquier valor superior a 0 indica que el atributo está presente en la entidad discursiva cuyo vector se está definiendo. El valor 0 se asigna a aquellos atributos que no pueden aplicarse a la entidad discursiva en cuestión. Por ejemplo, *María* tendrá el siguiente vector en los atributos de género: (neutro=0, masc=0, fem=3), y él tendrá el siguiente: (neutro=0, masc=3, fem=0). De este modo al hacer el producto escalar entre *él* y *María* dará como resultado 0. En la segunda regla, la de los atributos no restrictivos, estos atributos sólo tienen un valor posible que será asignado sólo si se cumplen las condiciones necesarias para su aplicación y que no será asignado en caso contrario. Además este valor será único y diferente para cada uno de los atributos. En el caso de los atributos sintácticos podrían considerarse los siguientes valores: (sujeto = 4, objeto directo = 3, objeto indirecto = 2, complementos = 1). De este modo en *La casa es grande*, el antecedente la casa, tendrá el valor 4 (sujeto).

De este modo los valores numéricos asignados al conjunto de atributos que definen una entidad discursiva concreta constituyen el vector que representa esa entidad. Y la comparación con otros posibles antecedentes se realizará mediante el producto escalar entre los vectores de cada uno de los antecedentes y el de la expresión anafórica. Como resultado se obtendrá una lista de las entidades ordenadas según su proximidad con el vector correspondiente a la expresión anafórica. Para mantener los resultados del producto escalar dentro de una escala fija se efectúa una normalización de las componentes del vector, con el objetivo de conseguir que el peso de los elementos que componen cada vector se pondere a una misma escala para todos los vectores. Para ello, se multiplica cada componente del vector por sc (Figura 12), de modo que el resultado de multiplicar el cuadrado de los elementos normalizados sea igual al cuadrado de una constante C, que se mantiene igual en todo el sistema. De este modo, el producto

escalar de dos vectores es siempre menor que 1, a menos que los dos vectores sean idénticos (será más próximo a 1 cuanto más parecidos sean los vectores).

$$sc = \sqrt{\frac{C^2}{\sum_{i=1}^{n} P_i^2}}$$

Figura 12. Constante a aplicar para la normalización de vectores.

En su implementación computacional, utiliza una gramática de cláusulas definidas, aprovechando el analizador sintáctico desarrollado por Amores [**Error! Reference source not found.**]. Para la identificación de unidades anafóricas recorre el enunciado de izquierda a derecha y anota en dos listas diferentes la aparición de una expresión anafórica o una entidad discursiva, cada una de ellas con su información sintáctica y morfológica.

Con referencia a la evaluación del sistema la principal crítica que surge está basada en la posible arbitrariedad a la hora de asignar manualmente estos valores. De todos modos, también apunta la posibilidad de probarlo con métodos conexionistas que asignen automáticamente estos valores en versiones posteriores. Este sistema resuelve manualmente los pronombres personales, reflexivos y las descripciones definidas, y computacionalmente sólo los pronombres personales y reflexivos. No resuelve computacionalmente las descripciones definidas porque necesitan fundamentalmente información semántica sobre su antecedente, y en muchas ocasiones conocimiento de inferencias sobre el mundo, concluyendo que el construir un módulo semántico que construya inferencias sobre conocimiento del mundo supone una gran complejidad. Respecto a sus limitaciones destaca que hay muchos tipos de anáfora por tratar, que no se ha implementado un módulo de gestión de la estructura discursiva y que no existe el ya mencionado mecanismo de generación de inferencias sobre conocimiento del mundo.

Una vez vistos varios ejemplos de sistemas integrados democráticos que utilizan dos distintos enfoques para el manejo de la información: o bien restricciones y preferencias, o bien solamente preferencias, es interesante comentar el trabajo de Mitkov [**Error! Reference source not found.**] en el que estudia la importancia del enfoque elegido para coordinar todas estas fuentes de información. Para ello comparará dos estrategias computacionales, cada una con diferentes estrategias de aplicación de estos factores[8]: una basada en restricciones y preferencias, y otra que utiliza únicamente preferencias sin descartar ningún antecedente asumiendo inicialmente que el antecedente que se examina es el correcto. La segunda estrategia a su vez utilizará la fórmula del razonamiento con incertidumbre (*uncertainty reasoning*) para aceptar o rechazar el antecedente en cuestión.

Ambas estrategias han sido estudiadas por Mitkov en diversos trabajos, la primera basada en restricciones y preferencias en [**Error! Reference source not found.**] y la segunda basada únicamente en preferencias en [**Error! Reference source not found.**]. En la primera se utiliza información morfológica, sintáctica, semántica y pragmática. Como restricciones escoge la concordancia, las restricciones c-dominio y la consistencia semántica. Y como preferencias aplica justo en este orden (los primeros criterios de preferencia serán los que más peso tengan en la elección final): la teoría del foco del discurso, paralelismo sintáctico, semántico y distancia entre la expresión anafórica y su antecedente.

En la segunda estrategia [**Error! Reference source not found.**] se utilizan técnicas de razonamiento con incertidumbre, justificando el uso de estas técnicas en la base de que en el proceso del lenguaje natural en muchas ocasiones partimos de información incompleta (por ejemplo en situaciones en las que no se ha entendido el texto completamente). Este sistema trabaja con los mismos factores que el sistema anterior pero sin distinguir entre restricciones y preferencias, asignando a cada factor un valor numérico que indique su aportación

[8] Mitkov nombra en sus trabajos a estas fuentes de información como factores.

a la identificación del antecedente. A este valor numérico le llamará factor de certeza (certainty factor o CF). Estos factores de certeza (CF) tendrán valores entre $0 < CF_t < 1$, en caso que se cumpla ese factor, y valores entre $-1 < CF_f \leq 0$, en caso que no se cumpla. Esto quiere decir, que si por ejemplo se cumple la concordancia entre la expresión anafórica y el antecedente se le asignará el valor CF_t, y en caso contrario se le asignará CF_f a este factor (concordancia). El proceso de identificación del antecedente emplea esta estrategia de razonamiento con incertidumbre: cada antecedente se evalúa como correcto o no según cada factor, asignándole su valor CF_t o CF_f según hemos visto en el anterior ejemplo. Progresivamente se va aplicando cada uno de los factores, calculándose la suma total de los valores CF de cada uno de ellos. Los antecedentes se prueban de derecha a izquierda (en primer lugar los que estén más cercanos a la expresión anafórica) mientras no se alcance uno cuya suma total de sus CF sobrepase un determinado valor umbral. La suma total de los CF se calcula por medio de la expresión aritmética mostrada en la Figura 13, en la que CF_1 representa el nuevo valor a partir del valor anterior CF_0 tras aplicar el factor s, el cual nos ha devuelto el valor CF_s. Por ejemplo, supongamos que un cierto antecedente ha alcanzado un valor CF_0 =0,5 tras aplicar algunos factores, si a continuación comprobásemos el factor restricción c-dominio y nos devolviese CF_s = 0,45, el valor final CF_1 = 0,5 + 0,45 – 0 * 0,45 = 0,725.

$$CF1\ (s, CF\ o) = CFs + CF\ o - CFs * CF\ o \quad \Leftrightarrow \quad CFs > 0\ ,\ CF\ o > 0$$

$$\left[1 - \min\left(|CFs|, |CF\ o|\right)\right] \quad \Leftrightarrow \quad CFs < 0 \quad \text{Ó} \quad CF\ o < 0$$

Figura 13. Esquema de cálculo del factor de certeza.

Ambas estrategias [**Error! Reference source not found.** y **Error! Reference source not found.**] utilizan los mismos factores: concordancia en género y número, paralelismo sintáctico, preferencia sobre las entidades que están en posiciones temáticas de la oración, consistencia semántica, paralelismo

semántico, preferencia por determinadas posiciones sintácticas (en primer lugar el sujeto de la oración anterior y después el objeto directo), preferencias inducidas por verbos o sustantivos, preferencia por sintagmas nominales que aparecen repetidos en el texto o que aparecen en las cabeceras de sección o capítulo y por último la distancia respecto la expresión anafórica.

La comparación de ambas estrategias se llevó a cabo sobre textos de informática, obteniendo un 83% de precisión para la basada en restricciones y preferencias, y un 82% para la basada en la técnica de razonamiento con incertidumbre con un CF_{umbral} = 0,7. La elección del valor umbral será un elemento determinante del éxito del sistema tal y como comprobó Mitkov, ya que al aumentarlo hasta 0,8, la precisión bajó hasta el 71%. Sobre estos resultados Mitkov saca las siguientes conclusiones: ambos métodos funcionan bien para la mayoría de los casos presentándose muy poca diferencia en cuanto a la precisión, aunque el primero parece ser más exacto pero menos seguro, poniendo como ejemplo las situaciones de excepción de la concordancia en número y género, es decir, situaciones en las que la expresión anafórica y su antecedente no concuerdan. Además, el segundo funciona mejor que el método basado en restricciones cuando se dispone de menos cantidad de conocimiento, aumentando su precisión cuantas más fuentes de información se emplean en la detección del antecedente.

Por último Mitkov sugiere una combinación de ambos métodos en una estrategia conjunta tal y como también propuso en [**Error! Reference source not found.**]. Esta consiste en que cada candidato se evalúa simultáneamente por ambos métodos, y en el momento que la respuesta de ambos coincida se detiene el proceso (acortando el proceso de búsqueda en comparación con el uso independiente de ambos métodos). De este modo, Mitkov considera que se ahorra un 10% del proceso de búsqueda en el caso del primer método, y un 33% en caso del segundo además de ganar en exactitud.

Como conclusión a su estudio, Mitkov considera que no es sólo importante el disponer de un conjunto adecuado de factores o fuentes de información lo que nos

lleva a un correcto tratamiento de la anáfora, sino que también es necesaria una clara estrategia de coordinación de estos factores. Continúa indicando que es difícil definirse por un sistema en el que encuentran restricciones o por otro en el que todo sean preferencias, pero sí teniendo claro que en el caso de utilizar el primero, hay que tener muy en cuenta las excepciones a esas restricciones (caso de la concordancia en número y género).

Respecto al orden de aplicación tanto de restricciones como de preferencias, para las primeras considera que el orden es indiferente puesto que cada una de las restricciones va a eliminar candidatos, mientras que para las preferencias sí que es importante puesto que algunas de esas preferencias van a entrar en conflicto en cuanto al antecedente seleccionado (casos de preferencia por antecedentes en la misma posición sintáctica y preferencias por el sujeto de la oración anterior). Del mismo modo, también considera que no todas las preferencias se pueden aplicar a cualquier tipo de texto, como por ejemplo una preferencia que propone Mitkov por la que se seleccionan los antecedentes que estén en las cabeceras de sección o capítulo, preferencia que evidentemente se aplicará en textos que estén organizados en forma de secciones o capítulos. En definitiva, considera que queda mucho trabajo por hacer en cuanto al estudio de la interrelación entre todas estas fuentes de información.

SISTEMAS POBRES EN CONOCIMIENTO

A continuación se van a describir diversos sistemas democráticos los cuales tienen en común la característica de intentar reducir la cantidad de información necesaria para la resolución de la anáfora pronominal. La cantidad de información que utilizan se limita prácticamente a la morfológica y sintáctica, eliminando la información semántica, inferencia de conocimiento e información del dominio o mundo exterior. El motivo de eliminar este tipo de información es fundamentalmente a que no se dispone habitualmente de la misma en aplicaciones reales de procesamiento del lenguaje natural, y también debido a que

aún sin incorporarla se obtiene una precisión suficientemente aceptable: por encima del 80%.

Una de estas primeras estrategias corresponde al algoritmo ya comentado anteriormente propuesto en Hobbs [**Error! Reference source not found.**], el cual utiliza únicamente información sintáctica y morfológica. Este algoritmo pone en práctica diversas restricciones sintácticas para las referencias pronominales. Para ello busca en el árbol sintáctico resultado del análisis de una oración para encontrar el antecedente de un pronombre. Una vez que encuentra el primer sintagma nominal que cumple estas restricciones, éste se acepta como antecedente. El orden que sigue para atravesar el árbol sintáctico es el siguiente: en primer lugar analiza la última oración que aparece en el texto y una vez que está dentro de ella recorre el árbol de izquierda a derecha y examina primero los nodos que están a un mismo nivel, para pasar después a un nivel más profundo (*breath-first*). Hobbs comprobó su algoritmo escogiendo 100 frases con diferentes ocurrencias de los pronombres *he*, *she*, *it*, y *they*, obteniendo un porcentaje de éxito del 81,8%.

Ya que este algoritmo trabaja únicamente sobre esta información sintáctica, sin tener en cuenta la información semántica y pragmática, resulta ser un algoritmo computacionalmente eficiente. Sin embargo, tiene la desventaja de estar muy limitado en el tratamiento de otros tipos de anáfora, y además no está implementado el módulo de análisis sintáctico que construya el árbol sintáctico correcto para todas las oraciones (este algoritmo no fue implementado, sino que los resultados fueron obtenidos tras un estudio manual).

En Lappin y Leass [**Error! Reference source not found.**] se hace la descripción de un algoritmo para la resolución de referencias pronominales que trabaja únicamente sobre información sintáctica con un alto porcentaje de análisis correctos: un 85%. Kennedy y Boguraev [93] proponen un algoritmo para la resolución de la anáfora pronominal que es una versión modificada del anterior trabajo de Lappin y Leass. En contraste con ese anterior trabajo este algoritmo no

necesita un análisis completo del texto. En su lugar trabaja sobre la salida de un etiquetador (*part-of-speech tagger o POS tagger*) enriquecido tan sólo con las funciones gramaticales de determinadas palabras. Al trabajar sobre la salida de un etiquetador presenta la ventaja de permitir su aplicación sobre sistemas de procesamiento del lenguaje natural que no puedan emplear componentes de análisis sintáctico robustos y fiables.

Este etiquetador devuelve un análisis muy simple de la estructura del texto: por cada palabra devuelve un conjunto de valores que indican sus características morfológicas, léxicas, gramaticales y sintácticas, dependiendo del contexto en el que aparece la palabra. A esta salida del etiquetador, su algoritmo añade la información correspondiente a la posición numérica de cada palabra en el texto, es decir, asigna un número secuencial a cada palabra en el texto. Esta información se utilizará para implementar las relaciones de precedencia.

La identificación de sintagmas nominales se obtiene a partir de reglas gramaticales que definen la composición de un sintagma nominal. Aparte de esta información gramatical se utilizará información referente al contexto en el que aparece el sintagma nominal, por ejemplo si se encuentra dentro de un sintagma preposicional o dentro de una oración de relativo. Además utilizará información concerniente al contexto de determinados pronombres, como podría ser el caso de *it*, e información como la del caso que apareciese como sujeto de un determinado grupo de verbos (*seem, appear*,…), información que nos pueda dar determinadas pistas sobre el posible antecedente a elegir.

También utiliza otra estructura de información que almacena las entidades del discurso, cada una de ellas representadas por información sobre ella misma y el contexto en el que aparece. Sin embargo, la única información con la que contará sobre su relación con otros referentes del discurso es en forma de precedencia respecto a la posición numérica que anota en el texto. La ausencia de información explícita sobre relaciones configuracionales marcan la diferencia fundamental entre este algoritmo y el de Lappin y Leass [**Error! Reference source not found.**]

en el que está basado. Este conjunto de entidades del discurso es el que utilizará el algoritmo de resolución de la anáfora.

En primer lugar este algoritmo activa el procedimiento de interpretación que realiza un análisis oración tras oración, interpretando los antecedentes del discurso en cada oración de izquierda a derecha. Cada nueva entidad del discurso presenta dos posibles interpretaciones: o bien como un nuevo participante en el discurso, o bien como una referencia a otra entidad previamente descrita en el texto.

La correferencia se determina al eliminar primero aquellos antecedentes a los que la expresión anafórica no se puede referir, para después aplicar unos criterios de preferencia sobre los antecedentes restantes. Los criterios de eliminación de antecedentes son básicamente dos. El primero basado en la información morfológica: concordancia en número, género y persona. Y el segundo basado en las restricciones c-dominio ya comentadas en la subsección que trata sobre la información sintáctica. Para implementar estas restricciones c-dominio, ya que no se tiene información configuracional, este algoritmo realiza inferencias de las funciones gramaticales según la precedencia entre constituyentes.

En relación a los criterios de preferencia a aplicar, Kennedy y Boguraev emplean diez condiciones basadas en información contextual, gramatical y sintáctica para determinar la elección de uno u otro candidato. Cada una de estas condiciones tiene asignada un valor numérico que se muestra en la Figura 14. La suma total de estos valores numéricos dará el valor total según el cual se ordenarán todos los candidatos para la correferencia. Este valor total lo denominan salience weight. Tal y como reseñan Kennedy y Boguraev, estos valores son totalmente arbitrarios, considerándose tan sólo importante la relación estructural entre ellos, es decir, lo que de verdad es importante es que SENT-S tenga el máximo valor y que el segundo máximo sea el correspondiente a CNTX-S, y así sucesivamente. Ellos justifican la relación estructural elegida por el estudio lingüístico y la experimentación realizada.

La suma total de estas preferencias para una determinada entidad no es fija durante la vida de la misma, sino que varía en función de su prominencia o reiteración en el discurso. De este modo, cuando un pronombre se refiere a una determinada entidad, su saliente *weight* se ve incrementado. En caso contrario, es decir, conforme pase el tiempo y la entidad no sea referida, este peso se verá disminuido hasta llegar a anularse.

Los candidatos son ordenados en función de este peso final, y el que tenga mayor peso es el que se determina antecedente del pronombre bajo consideración. En el caso de un empate, el antecedente más cercano al pronombre será el que se seleccione.

SENT-S: 100	**si esta en la misma oración**
CNTX-S: 50	**si está en el mismo contexto**
SUBJ-S: 80	**si su función es la del sujeto**
EXST-S: 70	**si está en una constricción existencial**
POSS-S: 65	**si GFUN= possessive**
ACC-S: 50	**si GFUN= direct objet**
DAT-S: 40	**si GFUN= indirect object**
OBLQ-S: 30	**si es complemento se una preposición**
HEAD-S: 80	**si EMBED= NIL**
ARG-S: 50	**si ADJUNCT= NIL**

Figura 14. Valores numéricos asignados a las preferencias.

En cuanto a la evaluación de este algoritmo, lo aplican exclusivamente para la anáfora pronominal con un porcentaje de éxito del 75%, que aunque es menor que el del algoritmo en el cual se basa, Kennedy y Boguraev lo justifican indicando que han trabajado sobre textos más variados y menos formales (el de Lappin y Leass [**Error! Reference source not found.**] conseguía un 85% con textos basados únicamente en manuales de informática).

Declerck [**Error! Reference source not found.**] propone un algoritmo el cual está basado en la unificación y que se lleva a cabo en la fase de análisis semántico.

Este algoritmo resuelve la anáfora pronominal interoracional. La operación de unificación es la que se utilizará para comparar la información contenida en el pronombre con la de los posibles antecedentes, y esta se aplicará en la fase de interpretación semántica de la frase, dentro del módulo de la gramática denominado de refinamiento (*refinement component of the gramma*r). El módulo de interpretación semántica está basado en el lenguaje de representación DPL (Dynamic Predicate Logic). Este lenguaje DPL es fruto de la investigación de la interpretación del lenguaje de cálculo de predicados de primer orden, considerándose como un primer paso para conseguir una teoría de la semántica del discurso. DPL está basado en la sintaxis de la lógica de predicados estándar, pero propone una nueva (dinámica) interpretación de los cuantificadores y conectivas que permiten el ligamiento de variables dentro y fuera de su ámbito, dependiendo de las interpretaciones de las correspondientes expresiones del lenguaje natural. De este modo considerará que los pronombres actuarán como variables, suposición no compartida por algunos trabajos como por ejemplo el la teoría de representación del discurso o el de Heim [**Error! Reference source not found.**].

El modo que DPL interpreta los distintos cuantificadores y conectivas es el que a continuación exponemos. Los cuantificadores existenciales y las conjunciones se consideran dinámicos externamente. Estos pueden ligar variables tanto dentro como fuera de su ámbito. Un ejemplo de este tipo de cuantificador sería el siguiente: *Un hombre$_i$ camina en el parque y él$_i$ silba. Él$_i$ es feliz*, en el que se mantendrá ligada la variable correspondiente al sintagma un hombre, fuera de su ámbito, unida con el pronombre él. El cuantificador universal y las implicaciones se considerarán dinámicas internamente, y sólo podrán ligar variables dentro de su ámbito. Por ejemplo, en: *Cada hombre$_i$ camina por el parque. Él$_j$ silba*, el pronombre él y su antecedente no compartirán la misma variable. La negación y la disyunción se considerarán como estáticas, por lo que no podrán ligar variables anafóricas. Por ejemplo, en la frase *Ningún hombre$_i$ camina por el parque. Él$_j$ silba*, o en la siguiente: *Un hombre$_i$ camina en el parque* o *él$_j$ silba*, nunca podrían

compartir la misma variable, es decir, nunca podrían correferir, el pronombre y el antecedente.

Stuckardt [**Error! Reference source not found.**] propone un algoritmo para la resolución de la anáfora pronominal (tanto para pronombres reflexivos como no reflexivos) y la anáfora producida por sintagmas nominales definidos. Este algoritmo está basado principalmente en las restricciones sintácticas derivadas del trabajo de Chomsky en su teoría de rección y ligamiento (Government and Binding Theory) [**Error! Reference source not found.**, **Error! Reference source not found.** y **Error! Reference source not found.**]. Stuckardt afirma que la implementación computacional de las restricciones sintácticas propuestas por Chomsky es inviable, ya que Chomsky las propone como un mero instrumento teórico. Chomsky asigna índices de modo aleatorio a cada sintagma nominal, y es durante la fase de interpretación semántica al traducirlos a su correspondiente fórmula lógica cuando los principios del ligamiento sirven como restricciones para filtrar la distribución de índices que se considerarán válidos para interpretar las correferencias. Una implementación directa de este procedimiento supondría una complejidad exponencial. La solución propuesta por Stuckardt supera esa complejidad exponencial, y también permite la interdependencia (*interdependency*) entre antecedentes. Esta interdependencia la interpreta que puede suceder fundamentalmente en tres situaciones. La primera cuando uno de los posibles antecedentes de un pronombre ya no está accesible, con lo cual se incumple el principio B de la teoría de Chomsky, como por ejemplo en la siguiente frase: *El barbero$_i$ contó al cliente$_j$ una historia mientras él$_i$ le$_i$ afeitaba*. La segunda corresponde a la situación en la que se ha de escoger entre antecedentes del discurso (fuera de la frase actual). Y la tercera sucede como consecuencia del ligamiento de cláusulas relativas.

El algoritmo aquí propuesto trabaja con una complejidad cúbica en su peor caso en función del número de sintagmas nominales, resolviendo aproximadamente un 90% de los pronombres aparecidos en textos sobre biografías de arquitectos. Este algoritmo consiste en los siguientes pasos:

- Por cada expresión anafórica Y se determinan el conjunto de posibles antecedentes X. Para ello se verifica la concordancia en número, género y persona entre Y y X. Además, en caso que el candidato X sea intraoracional se comprueban las restricciones sintácticas ya comentadas anteriormente.

- Por cada antecedente que quede en X, se le asigna un valor numérico que determina su admisibilidad (plausibility) como antecedente final. Este valor se basa en criterios ya comentados como el del paralelismo (preferencia de la expresión anafórica por aquellos candidatos que se encuentren en posiciones paralelas), proximidad, preferencia por el sujeto, etc. En función de este valor, se ordenarán todos estos antecedentes (de mayor a menor).

- También se ordenarán descendentemente las expresiones anafóricas en función del valor numérico de su mejor antecedente.

- Siguiendo el orden de expresiones anafóricas propuesto en el paso anterior, se irán asignando sucesivamente los mejores antecedentes de cada expresión anafórica siempre que no aparezcan problemas de interdependencia entre ellos. Por ejemplo, en la frase anterior del barbero, al determinar que el primer pronombre *él* refiere a *el barbero*, cuando posteriormente se resuelva el pronombre *le*, nunca se podrá correferir con el mismo antecedente.

Mitkov y Stys [**Error! Reference source not found.**] proponen otro 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 [**Error! Reference source not found.**] 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 antecedente de la expresión anafórica el que tenga mayor suma de estos valores.

Mitkov y Stys trabajan sobre la salida de un etiquetador gracias al cual obtienen la información léxica y morfológica. El motivo que reseñan por el que desechan la información semántica y pragmática es la propia complejidad de introducirlas en estrategias prácticas a la resolución de la anáfora, tanto desde el punto de vista humano como computacional. El conocimiento que utilizarán 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 acudirá a dos criterios para seleccionar el antecedente correcto: en primer lugar se escogerá el antecedente con mayor valor para el indicador de reiteración léxica, y en segundo lugar en caso que todavía persista el empate, se escogerá el más cercano.

Los indicadores de antecedente se aplicarán sobre los sintagmas nominales encontrados en una distancia máxima de dos oraciones respecto la expresión anafórica, y los utilizados por Mitkov y Stys son los siguientes:

- Se preferirán como antecedente los sintagmas nominales definidos (los que están modificados por un artículo, demostrativo o posesivo) antes que los indefinidos. Para ello asignarán valor 0 si el sintagma nominal es definido, y valor -1 si es indefinido.

- Son preferidos los sintagmas nominales que representan el tema (theme) del texto (información dada o conocida), en cuyo caso se les asigna valor 1 y en caso contrario valor 0. La nueva información que se añade al texto, o rheme proporciona información sobre el tema del texto. Aquí aplicará la siguiente regla heurística para determinar el tema del texto: se escogerá el primer sintagma nominal de cada oración como tema de la misma.

- Los sintagmas nominales que representan términos del dominio del texto son más probables de ser el antecedente correcto (valor 1, caso contrario 0).

- Si el verbo es uno de los siguientes: *discuss, present, illustrate, identify, summarise, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise*, *study, survey, deal o cover*, entonces se elige el primer sintagma nominal que los siga (valor 1 ó 0).

- Si el núcleo del sintagma nominal que precede al verbo es chapter, section o table, entonces se escogerá el que siga al verbo (valor 1 ó 0).

- Se preferirá un sintagma nominal si éste se repite varias veces. Se le asignará valor 2 si se repite dentro del mismo párrafo dos o más veces, 1 si se repite una vez y 0 en caso contrario.

- Si el sintagma nominal está en la cabecera de la sección y forma parte de la oración donde está la expresión anafórica, entonces se le asigna valor 1, caso contrario 0.

- Se escogerá el que ocupe la misma posición respecto al verbo que la ocupada por la expresión anafórica (valor 2 ó 0).

- Se preferirán en primer lugar los que estén en la cláusula previa (valor 2), después los que estén en la oración anterior (valor 1), después los que estén dos oraciones atrás (valor 0) y, finalmente, los que estén tres oraciones atrás (valor -1).

- Se le dará mayor preferencia (valor 0) a los sintagmas nominales que no formen parte de un sintagma preposicional (valor -1).

- Dada la estructura (*you*) V_1 NP_1 … conj (*you*) V_2 *it* (conj (*you*) V_3 *it*), se preferirá como antecedente del pronombre *it* a NP_1 (valor 1).

Este sistema 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 aplican obteniendo una precisión del 92,1%.

Stuckardt [**Error! Reference source not found.**] parte de la idea según la cual el conseguir un análisis sintáctico único es un objetivo poco acorde con el estado actual de desarrollo del campo del procesamiento del lenguaje natural a causa de la ambigüedad estructural (ligamiento de constituyentes como por ejemplo sintagmas preposicionales, adverbiales u oraciones de relativo; determinación de roles o funciones sintácticas como podrían ser la de sujeto u objeto directo; etc.). Por ello considera que un sistema que trate la anáfora ha de ser robusto en el sentido de ser capaz de superar esa inexactitud en el análisis sintáctico, o sea, que sea capaz de trabajar con estructuras sintácticas deficientes o incompletas. Stuckardt clasifica los sistemas robustos que tratan con esas estructuras deficientes en dos tipos:

- Modelos de descripciones superficiales (*shallow description model*), en los que se utilizan reglas heurísticas para reconstruir estas estructuras sintácticas. En estos casos el proceso de tratamiento de la anáfora se adapta al tipo de entrada de que dispone.

- Modelos de descripciones deficientes (*deficient description model*), los cuales extienden el tratamiento de la anáfora para trabajar con este tipo de estructuras, con lo que se utiliza la información sintáctica hasta donde es posible.

El trabajo de Kennedy y Boguraev [**Error! Reference source not found.**] comentado anteriormente lo encuadra dentro del primer tipo. También se comparan dos sistemas que trabajan con estructuras sintácticas incompletas o deficientes encuadrados en el segundo tipo, con el objetivo de aprovechar la salida del sistema de resolución de la anáfora para aplicarlo a una posterior fase

de desambiguación estructural. Ambos sistemas utilizan fragmentos lingüísticos para aplicar las restricciones basadas en la información sintáctica, deduciendo a partir de reglas heurísticas el resto de información configuracional no obtenida en la fase de análisis sintáctico. El primero de estos sistemas que utiliza reglas heurísticas se implementa para el alemán en el sistema de tratamiento de la anáfora descrito en un trabajo anterior del mismo Stuckardt [**Error! Reference source not found.**]. Con este sistema obtiene un 82% de precisión, en principio superior al 75% obtenido por Kennedy y Boguraev, aunque de nuevo no pudiéndose comparar ambos resultados al aplicarse sobre distintos textos e idiomas.

Baldwin [**Error! Reference source not found.**] presenta un sistema de resolución de la anáfora pronominal denominado CogNIAC con un 90% de precisión y un alcance (*recall*) superior al 60%. El sistema resuelve pronombres que no necesitan conocimiento del mundo ni un sofisticado tratamiento lingüístico. Lo que distingue a este sistema de los restantes es que únicamente resuelve pronombres de los que está completamente seguro, para los cuales no encuentra ninguna ambigüedad. Para ello exigirá que solamente quede un antecedente después de aplicar el algoritmo de resolución. Baldwin justifica esta afirmación indicando que hay frases definitivamente ambiguas, las cuales son también interesantes detectar, como por ejemplo la siguiente: *Pedro y Juan estaban trabajando juntos cuando de repente él se cayó*.

Este sistema utiliza la siguiente información: detección de oraciones, POS tagging, reconocimiento de sintagmas nominales (corregido manualmente), información del número y género, y árboles sintácticos parciales, es decir, elimina la información semántica y el conocimiento del mundo. Éste trabaja en el siguiente orden: en primer lugar se aplica el etiquetado palabra a palabra, para después reconocer sintagmas nominales simples (que no tengan anidados nuevos sintagmas nominales). Después identifica manualmente las cláusulas de cada oración, utilizando expresiones regulares para identificar sujeto, objetos y verbo. A continuación aplica una serie de restricciones (número, género y c-dominio) que

limitan el número de antecedentes sobre los que se aplicarán las siguientes preferencias heurísticas:

1. Único en el discurso: si hay un sólo antecedente se escoge éste como solución (evaluación: 8 correctos, 0 incorrectos).

2. Reflexivos: se escoge el antecedente más cercano de la oración actual en caso que se trate de un pronombre reflexivo (16 correctos, 1 incorrecto).

3. Único en la oración actual y previa: si hay un único antecedente en la oración actual y la previa, se escoge éste (114 correctos, 2 incorrectos).

4. Posesivos: si la expresión anafórica es un pronombre posesivo y hay uno con la misma estructura en la oración anterior, se consideran ambos correferenciales (4 correctos, 1 incorrecto).

5. Único en la oración actual: si sólo hay un antecedente en la oración actual se escoge éste (21 correctos, 1 incorrecto).

6. Antecedente y expresión anafórica son sujetos de la oración anterior y actual respectivamente, entonces se escoge este antecedente (11 correctos, 0 incorrecto).

CogNIAC resuelve los pronombres de izquierda a derecha en el texto. Para cada pronombre se aplican las anteriores reglas en el mismo orden que se han citado, de tal manera que si se cumple una de ellas no se continúa probando las demás. Si no se cumple ninguna de las reglas, entonces se deja sin resolver el pronombre.

También Baldwin compara su sistema con el algoritmo ya comentado anteriormente de Hobbs [**Error! Reference source not found.**]. Este algoritmo de Hobbs presenta una precisión del 81,6%, pero a diferencia del método propuesto por Baldwin, intenta resolver todos los pronombres sin detectar las situaciones de ambigüedad y además ordena todos los antecedentes (en el de Baldwin sólo se

escoge uno, sin evaluar los restantes). Baldwin aplicará ambos métodos sobre un texto narrativo en el que hay gran ambigüedad: una tercera persona cuenta la historia de otras dos personas del mismo sexo, y tratará únicamente pronombres singulares en tercera persona. Baldwin aplicará su método corrigiendo manualmente la detección de los sintagmas nominales con la intención de no degradar su sistema respecto al de Hobbs, el cual se ejecutaba manualmente. También evitará situaciones en las que se puedan arrastrar errores, es decir, en caso que al resolver un pronombre se produzca un error, lo corregirá antes de continuar el análisis. Con estas condiciones el algoritmo de Hobbs obtuvo un 78,8% de precisión y un 100% de alcance mientras que el de Baldwin obtuvo un 92% de precisión y un 64% de alcance. Baldwin también realizó pruebas ampliando el alcance de su sistema hasta un 100% para poderlo comparar con el de Hobbs, al añadir dos nuevas reglas a las anteriormente descritas:

7. Siguiendo la teoría del foco del discurso, la cual la implementó Baldwin en [**Error! Reference source not found.**], cada oración tiene un centro focal que mira hacia atrás (backward centre o Cb), el cual une el enunciado al discurso precedente, y supone la entidad focalizada en la oración. Si existe un antecedente en la cláusula actual que representa a este Cb, entonces se escoge éste como solución del pronombre.

8. Se escoge el antecedente más cercano a la expresión anafórica en caso que no se cumpla ninguna de las reglas anteriores.

Con estas nuevas reglas el algoritmo de Baldwin obtiene un 77,9% de precisión, o sea, incluso inferior al de Hobbs.

Además Baldwin presenta los resultados de la aplicación de CogNIAC a la extracción de información dentro de la evaluación del MUC-6 [**Error! Reference source not found.**]. Para ello se integró con otros módulos, como el analizador parcial desarrollado por Collins [**Error! Reference source not found.**], y también se modificó CogNIAC para adaptarlo al dominio sobre el que se trabajaría. Por

ejemplo, se eliminaron la cuarta y octava regla de preferencia, y se le añadieron módulos para detectar usos no anafóricos del pronombre *it*. Con ello obtuvo un 73% de precisión bastante inferior a otros resultados, pero atribuyendo este bajo rendimiento a los errores inducidos por el resto de módulos (mal etiquetado de palabras, errores de detección de sintagmas nominales, etc.).

SISTEMAS CONSULTIVOS: TEORÍA DEL FOCO DEL DISCURSO

El los anteriores apartados se han descrito sistemas integrados que hacen uso de un enfoque democrático para coordinar las distintas fuentes de información que intervienen en la resolución de la anáfora. A continuación se expondrán los sistemas que hacen uso del enfoque consultivo. El enfoque consultivo es característico debido a la selección de antecedentes gracias a la información recogida por una única fuente. Una vez que se ha propuesto un antecedente se consultarán otros tipos de conocimiento con el fin de recoger información suficiente para confirmar o rechazar dicho antecedente.

La fuente de información que normalmente hace la selección de antecedentes es este tipo de sistemas consultivos es la correspondiente a la estructura del discurso, y por lo tanto se suelen englobar epígrafe de sistemas basados en la teoría del foco del discurso, tal y como por ejemplo Ersan y Alkan [**Error! Reference source not found.**] nos definen en su trabajo. Ellos cuentan que los investigadores que trabajan en este tipo de sistemas intentan modelar la compleja estructura del discurso para que la anáfora, aceptada como un fenómeno a nivel del discurso, se resuelva utilizando esta estructura.

Esta teoría del foco del discurso está basada intuitivamente en la siguiente idea:

> Los participantes de un discurso coherente centran su atención sobre ciertas entidades del mismo. Algunas entidades se destacan sobre las demás conforme se desarrolla el discurso, por lo que ciertas expresiones anafóricas se utilizan como técnica para referirse a esas entidades. Para que se puedan resolver, el antecedente debe estar en la conciencia del

oyente; si esto no fuese así, el discurso estaría mal formado desde el punto de vista del oyente.

Según esta afirmación y suponiendo al discurso bien formado, si conseguimos obtener un modelo completo del discurso, entonces conseguiremos resolver estos tipos de anáfora. Por lo tanto, esta teoría se basa en este modelo del discurso y ha sido desarrollada inicialmente por Grosz [**Error! Reference source not found.** y **Error! Reference source not found.**], quien construye un modelo de focalización basado en tres mecanismos:

- Representación del foco explícito. Se utiliza para localizar los elementos de mayor relieve en el discurso, los cuales quedarán almacenados en los espacios focales.

- Representación del foco implícito. Contiene información sobre conceptos asociados con el elemento que está focalizado en ese momento.

- Cambio de foco. Su función es actualizar los elementos contenidos en los espacios focales a medida que avanza el discurso, es decir, el foco del discurso va a ser dinámico. Estos cambios se producen fundamentalmente de dos modos: explícita e implícitamente. El primer modo sucede cuando aparecen expresiones que nos avisan explícitamente del cambio de foco, tales como *hablemos sobre*…. El segundo modo corresponde a la propia estructura de la frase o del texto en general, y es su propia información la que nos indica el cambio de foco.

Estos mecanismos de focalización limitan la búsqueda a las entidades relacionadas con elementos focalizados, dejando en segundo término otras entidades menos accesibles. Esta teoría se aplica fundamentalmente a la resolución de referencias pronominales.

Sidner [**Error! Reference source not found.**, **Error! Reference source not found.** y **Error! Reference source not found.**] modifica la teoría del foco

esbozada por Grosz, al añadir las nociones de foco discursivo y foco agente que corresponden a entidades mencionadas en el enunciado que son focalizadas debido a la estructura sintáctica y a las relaciones temáticas de la oración. En estos trabajos, trata con referencias pronominales y sintagmas nominales definidos (sintagmas nominales precedidos por el artículo definido *the*, *this* o *that*), proponiendo diferentes algoritmos para cada tipo de anáfora. Las expresiones anafóricas se tratan de izquierda a derecha en la oración.

Por ejemplo, en *Juan sacó el perro a pasear*, el sintagma nominal Juan es el foco agente y el perro el foco discursivo. Sidner propone el siguiente algoritmo para la interpretación de la anáfora pronominal:

1) Si el pronombre está en una posición temática que no es la de agente, seleccionamos el foco discursivo como antecedente.

2) Si el pronombre está en una posición de agente, se tomará como antecedente el foco agente.

Podemos observar un ejemplo de la aplicación de este algoritmo sobre la siguiente frase: *Pedro y Juan compraron un libro*. Ellos suelen comprar uno cada semana. Aquí *Pedro* y *Juan* son el foco agente, mientras que un libro es el foco discursivo. De este modo *ellos* (posición de agente) referirá a *Pedro* y *Juan*. Igualmente, relacionaremos uno con el foco discursivo un libro.

Sidner reconoce la necesidad de aplicar restricciones sintácticas, semánticas y pragmáticas que puedan impedir la selección de un antecedente aunque la regla lo haya identificado como tal. Por ejemplo, en: *Pedro y Juan compraron un libro*. *Después, ellos se tomaron varios helados. Ellos pensaron que estos sabían realmente bien*, aquí el pronombre *estos* no puede relacionarse con un libro porque las restricciones sintácticas y semánticas lo impiden. Se selecciona entonces el siguiente foco discursivo, esto es, varios helados. De este modo, la entidad escogida como antecedente según la teoría del foco del discurso también ha de superar filtros semánticos y sintácticos, y no ha de provocar ninguna

contradicción en el mecanismo de inferencia. Los filtros sintácticos incluyen la concordancia morfológica y restricciones c-dominio. El mecanismo de inferencia lo aplica en ejemplos como el siguiente: *Llevé a mi perro al veterinario ayer. Él le mordió en la mano*, en el que elige al *perro* como antecedente del pronombre *él* puesto que los perros tienen mayor tendencia que las personas a morder, además de que no tienen manos.

En la primera frase del discurso, el algoritmo de Sidner intenta "adivinar" el posible foco del discurso, el cual lo ratificarán las oraciones posteriores. Para ello, utiliza una serie de reglas, como son el tipo de oración (por ejemplo, en el caso que se trate de una oración del tipo *is-a* o *there-is*, entonces escogerá el sujeto) o las relaciones temáticas del verbo. En las frases posteriores por cada expresión anafórica encontrada le aplica las restricciones sintácticas y semánticas, y los antecedentes resultantes los envía al mecanismo de inferencia, el cual confirmará o rechazará estos antecedentes.

Grosz, Joshi y Weinstein [**Error! Reference source not found.**] presentan un mecanismo de resolución de la anáfora que contrasta en algunos aspectos con el algoritmo de Sidner. La regla básica de interpretación anafórica está basada en un proceso de identificación de lo que denominan centros focales que vienen a sustituir a las nociones de foco discursivo y foco agente. Cada oración tendrá un centro focal que mira hacia atrás (*backward centre* o *Cb*) que une el enunciado al discurso precedente y supone la entidad focalizada en la oración. Y también cada oración dispondrá de un conjunto de centros focales que miran hacia delante (*forward centre* o *Cf*) que proporcionan un conjunto de entidades a las que pueden hacer referencia los siguientes enunciados. Este conjunto se encuentra ordenado por las propiedades gramaticales que se consideran importantes para obtener el grado de focalización en la frase.

Un elemento del conjunto *Cf* puede convertirse posteriormente en el discurso en un *Cb*. El proceso de identificación de antecedentes, al que denominan *centering*, consiste en comprobar si el centro del enunciado que se está procesando es el

mismo que el centro del enunciado anterior, en cuyo caso se podría haber empleado un pronombre.

Grosz y Sidner [**Error! Reference source not found.**] consiguen unir todas las teorías desarrolladas anteriormente, gracias a la definición de unos principios generales sobre la estructura del discurso, los cuales indican que la estructura básica del discurso consta de tres componentes:

- Estructura lingüística de la oración. El análisis lingüístico de la oración ayuda a localizar ciertos marcadores lingüísticos o *cue phrases* a partir de los cuales se produce un cambio de foco en el discurso.

- Estructura de intenciones. Determina la finalidad del discurso (*discourse purpose*). Todo discurso tiene una finalidad propia que lo diferencia de otros discursos y junto a esta, existen una serie de finalidades secundarias que se distribuyen a lo largo del discurso y contribuyen al cumplimiento de la finalidad principal.

- Estructura de atenciones. Es una abstracción del foco de atención de los participantes en el discurso a medida que éste se desarrolla. Como podemos suponer, se trata de una estructura dinámica que cambia y se actualiza durante el transcurso del discurso.

Con estos tres elementos construyen la estructura discursiva y la organizan en espacios focales y pilas donde se almacena la información. De este modo se restringe la búsqueda de antecedentes a los que estén dentro de los espacios focales.

El aspecto más importante de esta teoría es el de la ordenación de las entidades de los centros focales que miran hacia delante o *Cf*. Estas entidades se ordenan por criterios sintácticos, semánticos y léxicos. Los defensores de esta teoría afirman que esta trabaja correctamente para cualquier lenguaje siempre que se aplique una correcta ordenación del conjunto *Cf*. Por ejemplo, en [**Error!**

Reference source not found., **Error! Reference source not found.** y **Error! Reference source not found.**] se utiliza el siguiente criterio de ordenación de la lista de entidades *Cf* para el japonés: tópico > empatía > sujeto > objeto2 > objeto > otros, definiendo la empatía como la propiedad gramatical que indica la posición del hablante a la hora de describir la situación. En [**Error! Reference source not found.** y **Error! Reference source not found.**] se considera el siguiente criterio de ordenación de la lista *Cf* como el más adecuado para los lenguajes occidentales: sujeto > objeto1 > objeto2 > sintagmas preposicionales.

En la Figura 15 se modelizan los cambios de centro focal en función de los centros focales de una oración O_n y los de su oración previa O_{n-1}. Para ello se utiliza el concepto de centro preferido o Cp que consiste en el primer elemento de la lista Cf, y representará la predicción sobre el Cb (centro focal que mira hacia atrás) de la próxima oración.

	$Cb(0_n) = Cb\ (0_{n-1})$	$Cb(0_n) \neq Cb\ (0_{n-1})$
$Cb(0_n) = Cp\ (0_n)$	*Continuar*	*Cambio ligero*
$Cb(0_n) \neq Cp\ (0_n)$	*Retener*	*Cambio radical*

Figura 15. Cambios de foco del discurso.

Estos cambios de foco están guiados por las siguientes reglas:

- Si en una oración O_n hay un sólo pronombre, y este pronombre refiere algún elemento de $Cf\ (O_{n-1})$, entonces éste elemento constituirá el $Cb\ (O_n)$.

- Si en O_n no hay ningún pronombre o más de uno, el $Cb\ (O_n)$ será: à $Cb\ (O_{n-1})$ si $Cb\ (O_{n-1})$ aparece en O_n.

    1. En caso contrario, será el primer elemento de $Cf\ (O_{n-1})$ que aparece en O_n.

El orden habitual de preferencia en cuanto al cambio de foco sería el siguiente: continuar, retener, cambio ligero y cambio radical. Veamos un ejemplo que ayudará a entender estos cambios de foco:

- Inicialmente el texto comienza por la frase: *Juan es un buen chico*, para la cual tendremos Cb =Ø puesto que se trata de la primera oración del discurso, y como C*ƒ* = [*Juan*].

- Con la siguiente frase: *Él se encontró ayer con María*, tendremos Cb = *Juan* y C*ƒ* = [*Juan* > *María*].

- Consideremos a continuación las siguientes posibilidades:

    1. En caso que viniese después la frase: *A él le gusta ella*, quedaría Cb = *Juan* y C*ƒ* = [*Juan* > *María*], por lo que entraríamos en el caso de la Figura 15 denominado continuar.

    2. En caso que viniese: *A ella le gusta él*, tendríamos Cb = *Juan* y C*ƒ* = [*María* > *Juan*], entrando dentro del caso denominado retener.

    3. En el caso de: *Ella estaba con Lucía*, tendríamos Cb = *María* y C*ƒ* = [*María* > *Lucía*], es decir, dentro del caso cambio ligero.

    4. En caso de: *Lucía estaba con ella*, tendríamos Cb = *María* y C*ƒ* = [*Lucía* > *María*], dentro del cambio radical.

El algoritmo desarrollado por Brennan, Friedman y Pollard [18] construye y actualiza la estructura discursiva al mismo tiempo que ayuda a localizar los antecedentes anafóricos. Asimismo añade algunas extensiones al trabajo de Grosz, Joshi y Weinstein [74] para tratar mayor número de casos, por ejemplo construye una nueva regla que hace posible el cambio de foco discursivo en más ocasiones. Llevan sus ideas a la práctica con el desarrollo de un sistema que sirve de interfaz a una base de datos. Para ello hay dos procesadores, uno semántico que construye representaciones de las oraciones anotadas con información

sintáctica, información sobre concordancia e información sobre su función gramatical que pasa al otro procesador: el pragmático. Su algoritmo se puede resumir en tres fases:

- En la primera fase se localizan todas las expresiones anafóricas que se encuentren en la oración y se ordenan según su posición sintáctica con el sujeto de la oración en primer lugar. Con esta ordenación se establecen preferencias en la selección de candidatos. Además, se construyen todas las combinaciones posibles del centro focal que mira hacia atrás (Cb) y el que mira adelante (Cf).

- En la segunda fase se aplican una serie de filtros que eliminan como posibles antecedentes todas aquellas combinaciones de Cb y Cf que contradigan las reglas de indexación entre antecedente y expresión anafórica.

- Y finalmente en la tercera, se clasifica cada par (Cb, Cf) en función del tipo de transición que se da en el paso de una oración a otra. Así, por ejemplo, si el foco del discurso continúa siendo el mismo, los pronombres que aparezcan tendrán como antecedente este foco discursivo. Si por el contrario se detecta un cambio en el foco del discurso, tendrá preferencia como candidato a antecedente aquel par (Cb, Cf) que señale precisamente ese cambio y por tanto, las expresiones anafóricas que se hayan empleado considerarán el nuevo foco del discurso como antecedente preferido.

Por ejemplo, en el texto: *Brennan drives an Alfa Romero. She drives too fast. Friedman races her on weekends. She often beats her*, en la primera oración el Cb y Cf coinciden en Brennan. En la segunda oración el Cb y Cf siguen siendo los mismos por lo que el foco del discurso sigue siendo *Brennan*. Como no se ha producido ningún cambio en el foco, el pronombre she tiene como antecedente la entidad *Brennan*. En la tercera, el Cb es *Brennan* pero el Cf cambia a *Friedman*,

esto es, hay un cambio de foco que influye en la selección del antecedente de *she* en la cuarta oración y hace que *Friedman* desplace a *Brennan*.

El inconveniente de este sistema es que no tiene en cuenta la incorporación de otras fuentes de conocimiento tales como la información semántica, el conocimiento del mundo o restricciones sintácticas que podrían rechazar el antecedente preferido por el foco discursivo.

Carter [**Error! Reference source not found.** y **Error! Reference source not found.**] desarrolla el sistema SPAR (*Shallow Processing Anaphor Resolver*) con el deseo de comprobar la hipótesis de que no es necesario efectuar un tratamiento demasiado profundo del lenguaje para obtener resultados correctos en la resolución de ambigüedades. Carter afirma que incluso es posible limitar la cantidad de conocimiento del mundo necesaria para tratar computacionalmente el lenguaje, siempre que la información estrictamente lingüística esté suficientemente fundada. En este sistema los candidatos a antecedente son seleccionados tras la aplicación de las reglas contenidas en el módulo de gestión del discurso. Este sistema se compone de los siguientes módulos: sintáctico, semántico, módulo con reglas de gestión y actualización de la estructura del foco del discurso (módulo dominante) y un módulo que aplica reglas de inferencia sobre el mundo. El algoritmo utilizado se puede sintetizar en los siguientes pasos:

- Aplicación de los módulos sintáctico y semántico.

- Aplicación del módulo de gestión discursiva basado en reglas sobre la focalización de las entidades discursivas. Éste módulo está basado en las reglas de Sidner [**Error! Reference source not found.**] pero añade un nuevo concepto: *la preferencia débil*. Este concepto se utilizará para determinar situaciones en las que no se tenga clara la preferencia, en cuyo caso el resto de módulos del SPAR tomarán la decisión final sobre qué candidato es el correcto. Un ejemplo sería la diferente selección del antecedente en función del verbo que se utilice en la segunda oración: *The*

monkey picked a banana$_i$. *The elephant ate it*$_i$, o para el verbo *attacked*: *The monkey*$_i$ *picked a banana*. *The elephant attacked it*$_i$. De este modo, si se hubiese permitido una preferencia fuerte sobre uno de los dos candidatos, *monkey* o *banana*, siempre existiría la posibilidad de que el candidato propuesto fuera el erróneo.

- Aplicación de reglas de concordancia sintáctica y restricciones semánticas
- Aplicación de reglas de restricción sintáctica para la localización del antecedente de un pronombre.
- Reglas de inferencia que se aplicarán sólo cuando todavía existen relaciones anafóricas ambiguas tras la fase anterior.
- Aplicación de reglas heurísticas si todavía hay ambigüedad.

Los resultados que obtiene son satisfactorios ya que Carter obtiene un porcentaje del 93% en la identificación de los antecedentes correspondientes a 242 pronombres.

Luperfoy [**Error! Reference source not found.**] construye una representación del discurso en tres capas y la aplica a un interfaz de diálogos en lenguaje natural como parte del proyecto *Human Interface Tool Suite (HITS)* del *MCC Human Interface Laboratory*. Sus aplicaciones son un editor del conocimiento para la base de conocimiento *Cyc* [**Error! Reference source not found.**], un editor de iconos para el diseño de pantallas para máquinas fotocopiadoras y una herramienta de recuperación de información. En su sistema Luperfoy trata con sintagmas nominales definidos.

Las tres capas de las que habla son la capa lingüística, el modelo del discurso y la base de conocimiento. En la primera capa se introduce un objeto lingüístico por cada expresión anafórica y posible antecedente que aparezca en el texto. En la segunda se crea otro objeto por cada concepto que se desarrolle en el discurso. Y finalmente, en la tercera capa, la base de conocimiento, representa el sistema de

conocimiento de una persona que participe en el diálogo. Luperfoy considera muy importante la separación entre la segunda y tercera capa para distinguir entre la comprensión del discurso y los conocimientos previos al propio discurso. Es decir, en algunas ocasiones aparece un objeto en el discurso que no es familiar al oyente, por lo que momentáneamente no se puede enlazar en la base de conocimiento. En estos casos se crearía un objeto del discurso sin especificar, quedando a la espera de que el discurso se desarrolle y nos ofrezca nueva información que nos permita clarificar su enlace en la base de conocimiento.

Hardt [**Error! Reference source not found.**] propone un interesante sistema basado en esta teoría del foco del discurso para tratar con las denominadas identificaciones descuidadas (*sloppy identity*) producidas en la elipsis verbal y para tratar con los pronombres del tipo "cheque" (*paycheck pronouns* conocidos así por el ejemplo de Kartunnen: *The man who gave [his paycheck]$_1$ to his wife was wiser than the one who gave it$_1$ to his mistress*). En estos casos la expresión anafórica y su antecedente no tienen idéntico significado. Por ejemplo, en: *Pedro quiere a su gato. Juan también*, al resolver la elipsis verbal nos quedaría en la segunda frase: *Juan también quiere a su gato*, con lo que la referencia producida por el posesivo su habría que resolverla nuevamente (¿Juan quiere a su gato, o al gato de Pedro?). Basándose en el sistema desarrollado por Muskens [**Error! Reference source not found.**], Hardt se propone plantear referencias al foco del discurso de manera explícita, para que después de resolver la elipsis el objeto elidido siga haciendo referencia al nuevo foco del discurso de la oración. Es decir, sobre el anterior ejemplo, en la primera frase *Pedro* es el foco del discurso, pero en la segunda frase *Juan* pasa a ser el nuevo foco. De este modo, la entidad su gato en la primera frase referirá a su foco (*Pedro*), y al recuperarse en la segunda frase pasará a referir al nuevo foco de la misma (*Juan*), con lo que no habría que repetir el proceso de resolución de la anáfora.

Una de las desventajas de la teoría del foco del discurso es la comentada por Manabu y Kouji [**Error! Reference source not found.**], según la cual esta teoría no ha sido totalmente probada en textos reales no restringidos, sino únicamente

en sucesivas oraciones simples que habitualmente contienen un sólo verbo. Manaby y Kouji proponen precisamente dos métodos de manejar oraciones complejas en las que haya más de un verbo: *Oración1 Conjunción Oración2*, donde *Oración1 y Oración2* han de ser oraciones simples. El primer método de tratar estas oraciones sería el considerar la conjunción como si fuese un punto de separación entre frases (con Cb y Cf independientes entre sí). Y el segundo método consistiría en tratar toda la oración como un conjunto, es decir, no considerarlas como frases independientes (con un único Cb y Cf). Ellos comparan ambos métodos concluyendo que el segundo reporta ciertos problemas con lo que consideran preferible tomar las oraciones coordinadas como oraciones simples. Esta misma solución es seguida en otros trabajos como por ejemplo los de Di Eugenio [**Error! Reference source not found.**] y Kameyama [**Error! Reference source not found.**].

Azzam [**Error! Reference source not found.**] también intenta solucionar la anterior limitación de la teoría del foco del discurso sobre oraciones complejas, proponiendo un algoritmo para la resolución de la anáfora pronominal intrasentencial que ocurre en oraciones que contienen cláusulas subordinadas. Este algoritmo toma como base los trabajos de Sidner [**Error! Reference source not found.**, **Error! Reference source not found.** y **Error! Reference source not found.**], intentando expandirlos para el tratamiento de este tipo de oraciones. Según Azzam, el algoritmo propuesto por Sidner debe ser mejorado en la anáfora intrasentencial, ya que éste sólo trabaja con los antecedentes de las oraciones anteriores, sin tener en cuenta los que aparecen en la oración actual (opinión también compartida por Carter [**Error! Reference source not found.**]). Otro problema que encuentra en el algoritmo de Sidner es cuando la primera oración del texto presenta una expresión anafórica (normalmente en una oración subordinada), con lo que el algoritmo que Sidner propone para seleccionar ese foco inicial no funciona correctamente. En la solución propuesta por Azzam, se considera que las oraciones que contienen cláusulas subordinadas incluyen varias situaciones o hechos elementales. Por ejemplo, la frase: *[Tres de las empresas*

más importantes del mundo]$_i$ dijeron que ellas$_i$ están formando una plataforma común, es dividida en dos hechos, por una parte, el hecho de que alguien está diciendo "algo", y por otra parte lo que se está diciendo. Cada uno de estos hechos será el marco sobre el que se aplicará la teoría de foco del discurso propuesta por Azzam, igual que si se tratase de sucesivas oraciones simples. De este modo se evita el problema ocasionado por las expresiones anafóricas encontradas en una cláusula subordinada de la primera oración. Para los pronombres posesivos, recíprocos y reflexivos que aparezcan en el primer hecho del texto prevé un tratamiento especial. Por ejemplo, en: *Las empresas son por sí mismas*…, en primer lugar intenta resolver estas referencias antes de aplicar el algoritmo de adivinar el foco de esta primera oración, suponiendo que no habrá gran ambigüedad en esta etapa del discurso. El proceso de división de la frase en hechos se lleva a cabo por medio de un analizador conceptual implementado en [**Error! Reference source not found.** y **Error! Reference source not found.**] por el mismo Azzam.

Strube [**Error! Reference source not found.** y **Error! Reference source not found.**] también extiende esta teoría del foco del discurso para la resolución de la anáfora intrasentencial y para el manejo de oraciones complejas en alemán. Aquí supone para las oraciones complejas un Cb y un Cf por cada cláusula, considerando inadecuada para el alemán la ordenación de la lista de entidades Cf basado en los criterios de roles temáticos, tal y como Suri y McCoy [**Error! Reference source not found.**] proponen en, donde además se plantea una estrategia mixta para el tratamiento de las oraciones complejas, según la cual en determinadas ocasiones se preferirán antecedentes intraoracionales y en otras ocasiones interoracionales. En lugar de ello, considerará más adecuada una ordenación en función de la información estructural de la oración.

Tal y como cuenta Di Eugenio [**Error! Reference source not found.**], otro tema no resuelto todavía en esta teoría del foco del discurso es el del tratamiento de los posesivos y el determinar cómo afectan éstos a los centros focales. Di Eugenio llega a la conclusión de que Cb no parece ser afectado por estos posesivos,

mientras que Cf sí que necesita ser modificado. Según Di Eugenio, un sintagma nominal de tipo posesivo se refiere a dos entidades: el poseedor (P_{or}) y lo poseído (P_{do}). Lo poseído se corresponde con el sintagma nominal completo, por lo que su posición dentro de Cf es determinando por su función gramatical. Con respecto al poseedor, considera (tras pruebas heurísticas) que la posición que le debe corresponder es la inmediatamente anterior al objeto poseído si éste es inanimado, y como siguiente si es animado.

Ersan y Akman [**Error! Reference source not found.**] muestran la ineficacia de esta teoría del foco del discurso ante los casos de anáfora que se suceden en estructuras paralelas. Por ejemplo, en el siguiente texto: *The green Whitierleaf is most commonly found near the wild rose. The wild violet is found near it too*, el pronombre *it* se refiere al sintagma nominal *the wild rose* al utilizar la noción de paralelismo entre estructuras sintácticas, sin embargo, al aplicar la teoría del foco, esta elegiría como antecedente a *Whitierleaf*.

Mitkov [**Error! Reference source not found.** y **Error! Reference source not found.**] intenta combinar los métodos lingüísticos clásicos con la teoría del foco del discurso. Para ello plantea la posibilidad de aplicar la información morfológica, lingüística y semántica del modo estudiado anteriormente, junto con esta teoría aplicada como preferencia en caso de ambigüedad a la hora de seleccionar un antecedente. Además, Mitkov realiza la aportación de aplicar un método probabilístico a la selección del foco del discurso en la primera frase del texto. Este método lo lleva a cabo en dos pasos: selección de una serie de reglas heurísticas obtenidas de un estudio previo del dominio del texto, las cuales nos indican modelos seguidos por el foco del discurso durante el texto analizado, y como segundo paso, el cálculo de la probabilidad de un determinado segmento a ser el foco en la primera frase en función de las reglas anteriores.

3.1.2.3 Sistemas alternativos

Anteriormente se han revisado sistemas integrados en lo cuales no se a utilizado la información obtenida de corpus. Enseguida se van agregar unos sistemas, los cuales emplean principalmente información de este tipo obtenida del estudio del corpus sobre el cual se el que posteriormente se evaluará el sistema que se propone.

Un sistema alternativo es el que propone Dagan e Itai [**Error! Reference source not found.**]. Ellos desarrollaron una estrategia estadistica con la cual lograron la desambiguación de los pronombres. Para esto aplicaron un método sobre ocurrencias del pronombre *it* el cual se encuentra en una serie de oraciones seleccionadas aleatoriamente. Hacian uso unicamente de la información de las plantillas obtenidas del análisis previo del corpus. Dichas plantillas consiten en patrones asociados a casa tipo de expresión anafórica con la posición que ocupaba su antecedente. Después de nalizar previamente el corpus, estos patrones se aplican a las nuevas expresiones anafóricas que se encuentren, seleccionando el antecedente que se encuentre en la misma posición que muestra la plantilla. La evaluación del método obtuvo un 87% de éxito, eso sí, eliminando manualmente los casos de pronombres *it* no anafóricos.

Un ejemplo de la aplicación de este método podría ser el siguiente: *They know full well that the companies held tax money$_i$ aside for collection later on the basis that the government$_j$ said it$_j$ was going to collect it$_i$*. Tenemos dos ocurrencias del pronombre *it*: una funcionando como sujeto y otra como objeto, ambas del verbo *collect*. Existen tres posibles antecedentes: *money*, *collection* y *government*, y después del análisis del corpus se dispone de las plantillas mostradas en la Figura 16 que nos miden el número de ocurrencias de cada antecedente funcionando como sujeto o como objeto del verbo *collect*. Según estas estadísticas concluyen que *government* es el antecedente del primer pronombre *it*, y que *money* es el antecedente del segundo.

Sujeto	Verbo	Número de ocurrencias en el texto
collection	*collect*	*0*
money	*collect*	*5*
government	*collect*	*198*
Verbo	**Objeto**	**Número de ocurrencias en el texto**
collect	*collection*	*0*
collect	*money*	*149*
collect	government	0

Figura 16. Plantillas obtenidas del análisis del corpus según Dagan e Itai.

Nasukawa [**Error! Reference source not found.**] él hace la descripción de una estrategia muy simple la cual utiliza información interoracional extraída de un texto fuente para mejorar la precisión de su sistema de resolución de la anáfora pronominal. Una cualidad de este sistema es que no necesita información del mundo exterior, ni hace un análisis sintáctico, morfológico o semántico. Nasukawa nos sugiere que las plantillas que definen la relación entre modificadores y modificados pueden resultar útiles para la detección del antecedente correcto de un pronombre. Por ejemplo, de la oración *He moved his residence* extrae la plantilla por la que *residence* puede ser objeto del verbo *move*. Del mismo modo también considera que la reiteración puede considerarse como otro indicador de la presencia del antecedente, expresada esta reiteración por la sucesión de sintagmas nominales con la misma palabra núcleo. Igualmente plantea la incorporación de otras reglas heurísticas dependientes del dominio en la resolución, como por ejemplo la preferencia por el sintagma nominal con la función de sujeto en lugar de los que ocupen la función de objeto.

El sistema de Nasukawa obtuvo una precisión del 93,8%, cuando lo probó sobre 1904 oraciones consecutivas de ocho capítulos de dos manuales diferentes de informática, que contenían 112 pronombres en tercera persona.

Connolly, Burger y Day [**Error! Reference source not found.**] hicieron la descripción de una estrategia al tratamiento de la anáfora la cual estaba basada en el aprendizaje computacional. Ellos enfocan la anáfora como un problema de clasificación, es decir, dada una expresión anafórica y dos posibles antecedentes reducen el problema a la elección entre una de esas dos clases (cada uno de los dos antecedentes). Una vez elegida una de las clases se compara este antecedente con otro, repitiendo el proceso sucesivamente hasta que no quede más que una solución posible. Para elegir entre cada pareja de antecedentes se codifica la información de forma discreta como vectores de atributos (de modo similar al trabajo de Rico [**Error! Reference source not found.**]), donde cada atributo describe propiedades de la expresión anafórica, los antecedentes y la relación entre ellos tres.

Rocha [**Error! Reference source not found.**] nos describe una estrategia para el tratamiento de la anáfora en diálogos, haciendo uso de conversaciones en inglés y portugués. Para ello utiliza la información probabilística obtenida de la anotación manual de un corpus de conversaciones de este estilo (sobre unos 3000 casos de anáfora para cada lenguaje). Esta información es almacenada en un modelo basado en árboles probabilísticos. Según el estudio que realiza del corpus, se anotarán los casos de anáfora de acuerdo a cuatro propiedades. La anotación se realizará manualmente por analistas con el objetivo de marcar las entidades importantes en el discurso: tópico del discurso tanto global como local, expresiones anafóricas y sus antecedentes. Para cada tipo de expresión anafórica se obtendrá su árbol probabilístico.

Cada expresión anafórica se anota en el corpus de acuerdo a las siguientes cuatro propiedades:

- *Tipo de expresión anafórica*, distinguiendo a su vez entre las expresiones formadas por palabras (pronombres con función de sujeto, verbos anafóricos o adverbios), por sintagmas (sintagmas nominales definidos) y casos de anáfora de tipo “one”.

- *Tipo de antecedente*, según sea éste explícito, implícito (no aparece en el texto) o no referencial (casos de pronombres no anafóricos como el pronombre *it* cuando actúa como sujeto de una oración impersonal, por ejemplo en la frase: *It's raining*). En el caso que el antecedente aparezca explícitamente en el texto, se le añade a esta propiedad el número de antecedente según una lista de antecedentes posibles aparecidos en el texto.

- *Grado de topicalidad del antecedente*. En caso que sea el tópico del segmento en el que aparece se le clasifica como elemento temático. Si además éste tiene relación con los participantes del diálogo o con algún elemento del tópico global del discurso, entonces se le clasifica como *elemento temático del discurso*.

- *Información psicolingüística*, en la que se intenta codificar observaciones encontradas en el análisis del corpus que demuestran que para expresiones anafóricas de la misma categoría se pueden necesitar diferentes estrategias para su resolución. Estas estrategias las codificarán en cuatro grupos: basadas en procesos léxicos (utilizando conocimiento semántico asociado a la expresión anafórica, conocimiento del mundo y reiteración léxica), basadas en el discurso, según la posición (nociones de paralelismo entre expresión anafórica y su antecedente) y según procesos sintácticos.

Para cada una de esas cuatro propiedades se obtendrán datos probabilísticos de la distribución de cada una de sus subcategorías en los casos de anáfora estudiados.

En este tipo de sistemas alternativos es importante mencionar los trabajos explicados anteriormente de Mitkov [**Error! Reference source not found.** y **Error! Reference source not found.**] en lo que hace uso de una serie de indicadores de antecedente que consisten en reglas heurísticas de preferencia obtenidas del

estudio del corpus sobre el que se aplicarán posteriormente dichas reglas. En definitiva, se trata de información obtenida a partir del estudio del corpus, o sea, reglas dependientes del dominio sobre el que se trabajará posteriormente. Por ahora tales reglas se obtienen de forma manual e indican comportamientos usuales que siguen las expresiones anafóricas para localizar sus antecedentes.

CAPITULO 4 SISTEMA PARA LA RESOLUCIÓN DE LA ANÁFORA

En este capítulo se describe el sistema para la resolución de la anáfora.

4.1 El método implementado

El sistema desarrollado implementa una variante del algoritmo MARS propuesto por Mitkov [**Error! Reference source not found.**] que será descrito a continuación. Posteriormente se mencionarán las variaciones a este método que hice para adaptarlo para el idioma español.

MARS opera en cinco fases. En la *fase 1*, el texto a procesar es parseado sintácticamente usando el parser FDG [1] que devuelve “partes de oración” (parts of speech), lemas morfológicos, funciones sintácticas, número gramatical y las relaciones de dependencia entre los elementos (*tokens*) del texto que facilitan la extracción de frases nominales complejas.

En la *fase 2*, se identifican los pronombres anafóricos. En esta implementación MARS solo está enfocado a resolver pronombres en tercera persona y posesivos en plural y singular que demuestren anáfora nominal de identidad de referencia (*identity-of-reference nominal anaphora*).

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 estarán sujetos 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 (un total de 10) al conjunto de candidatos contendientes. Cada factor aplica una ganancia numérica a cada candidato.

Lo definido (definiteness)

Los sustantivos "*definidos*" de las oraciones previas tienen más posibilidad de antecedencia 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.

Lo dado (givenness)

Los sustantivos de las oraciones previas que representan "la información dada" (tema)[1] se consideran buenos candidatos a antecedentes y obtienen una puntuación de 1 (los candidatos que no representan el tema obtienen 0).

Verbos indicativos (Indicating verbs)

Si un verbo es miembro del siguiente conjunto de verbos {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 (puntuaciones 1 y 0). La evidencia empírica sugiere que debido a la relevancia de los sustantivos que los siguen, los verbos mencionados son buenos indicadores.

---

[1] Se utilizó la heurística de que la información dada es la primera frase nominal en una oración no imperativa.

Reiteración léxica (Lexical reiteration)

Los sustantivos léxicamente reiterados son buenos candidatos a antecedentes (una frase nominal obtiene una puntuación de 2 si en el mismo párrafo es repetida 2 o más veces, 1 si es repetida una vez y 0 si no se repite). Los elementos léxicamente reiterados incluyen frases nominales sinonímicas que a menudo pueden estar precedidas por artículos *definidos* o 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*").

Frases nominales "no preposicionales" ("Non-prepositional" noun phrases)

Una frase nominal "*no preposicional pura*" tiene mayor preferencia que una frase nominal que forma parte de una frase preposicional (0, -1). Ejemplo:

Pon el libro$_i$ en el estante, asegúrate que los niños puedan alcanzarlo$_i$.

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*).

Preferencia de patrón de colocación (Collocation pattern preference)

Esta preferencia se le da a los candidatos que tienen un patrón de colocación idéntico a un pronombre (2, 0). La preferencia de colocación está restringida a los patrones "frase nominal (pronombre), verbo" y "verbo, frase nominal (pronombre)".

Referencia inmediata (Immediate reference)

En manuales técnicos, la pista de "la referencia inmediata" puede ser muy útil en la identificación del antecedente. La heurística usada es que en contrucciones del tipo "… (tú) V_1 NP … con (tú) V_2 lo/la (con (tú) V_3 lo/la)", donde con $\in$ {y/o/antes de/ después de …}, la frase nominal después de V_1 tiene mucha posibilidad de ser

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

Para encender la impresora, presione el botón$_i$ y manténgalo$_i$ presionado por un momento.

Desengrape el papel$_i$, acomódelo$_i$, después cárguelo$_i$ en la bandeja.

Distancia referencial (Referential distance)

En oraciones complejas, las frases nominales de la sección anterior son los mejores candidatos a antecedentes de una anáfora en la sección precedente, seguidos por las frases nominales de la oración anterior, después por sustantivos situados 2 oraciones más atrás y finalmente por sustantivos 3 oraciones atrás (2, 1, 0, -1).

Preferencia de términos (Term preference)

Las frases nominales que representan términos del tema, tienen más posibilidad de ser antecedentes que sustantivos que no lo son (1, 0)

En la *fase 5*, el candidato con el marcador más elevado es seleccionado como el antecedente del pronombre. Los lazos 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 9 parámetros, cada uno representando un indicador de antecedencia

$$score_k = \sum_{i=1}^{i=9} x_{k_i}$$

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 mencionarán las variaciones al método original.

En la *fase 1*, para el caso de nuestro sistema, se utilizó un parser desarrollado en el Laboratorio de Lenguaje Natural del Centro de Investigación en Computación del Instituto Politécnico Nacional.

En la *fase 2*, se identifican los pronombres anafóricos. Esta implementación se enfoca a resolver pronombres personales en tercera persona.

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: *definiteness, givenness*, v*erbos indicativos (indicating verbs)*, *reiteración léxica (lexical reiteration)*, *frases nominales "no preposicionales" ("non-prepositional" noun phrases)*, *Preferencia de patrón de colocación (Collocation pattern preference), distancia referencial (referential distance)*, contemplados en el método original.

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, 0).

4.2 Uso del parser del español

El programa denominado «PARSER» permite investigar la estructura sintáctica y morfológica de oraciones en español y proporciona la información detallada sobre el comportamiento interno de los componentes del analizador. Con este programa

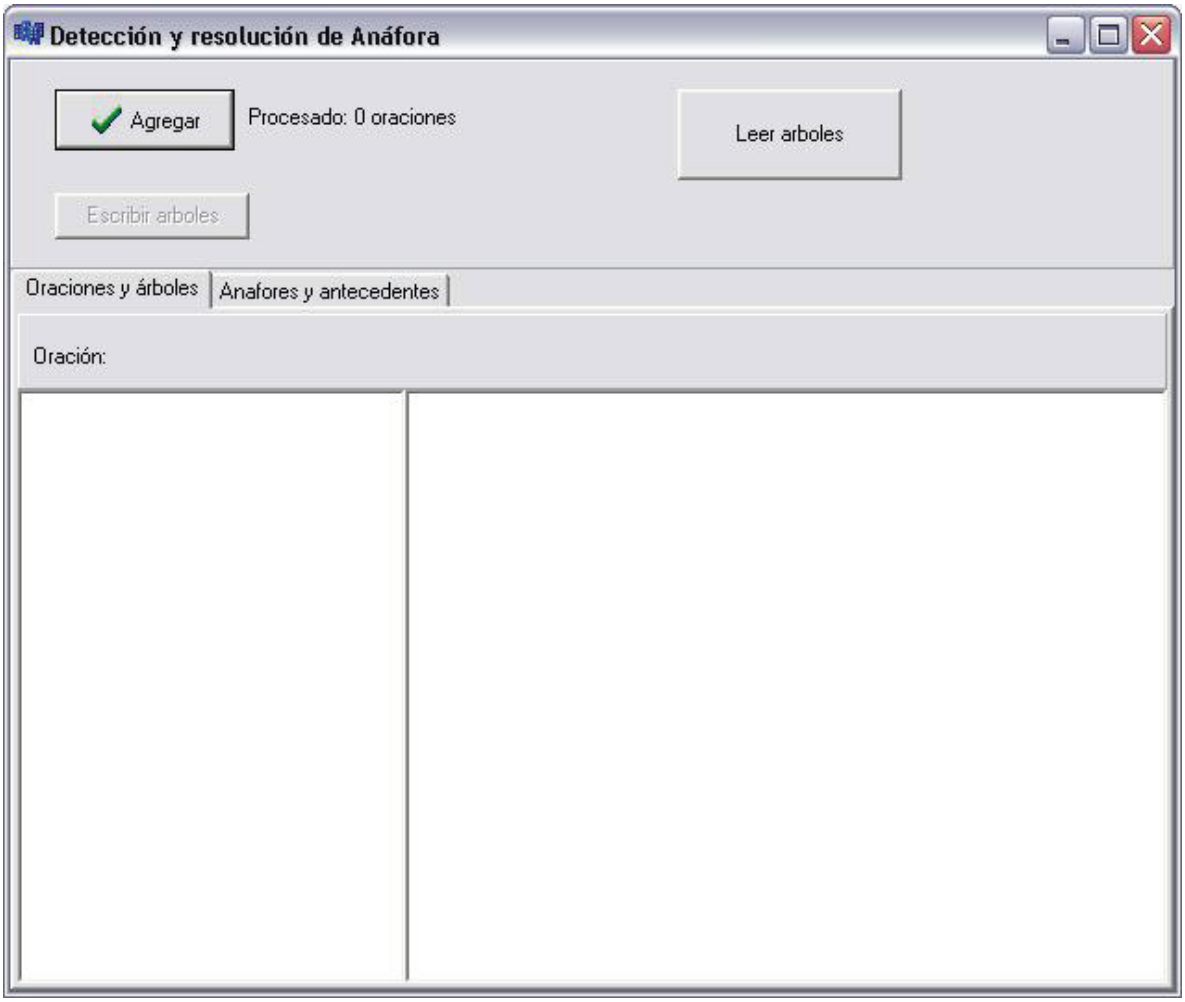
se puede aprender el formalismo de gramáticas independientes del contexto. También es posible desarrollar y probar este tipo de gramáticas.

El núcleo del sistema es 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.

4.3 Descripción del sistema

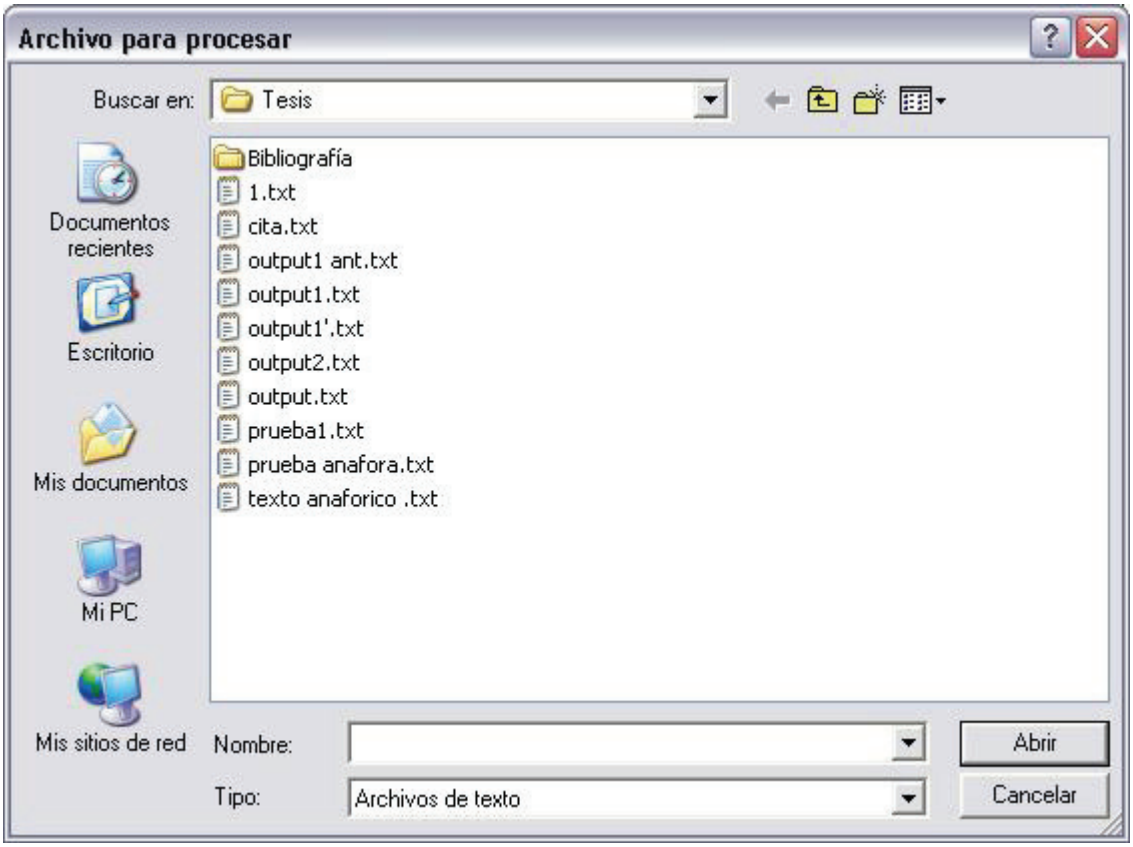
El sistema fue desarrollado en el leguaje de programación C++ Builder 6.0. En esta sección se describirá la interfaz gráfica con la que interactúa el usuario.

Al iniciar el sistema, se nos muestra la pantalla principal:



En esta parte del sistema, podremos presionar el botón Agregar,

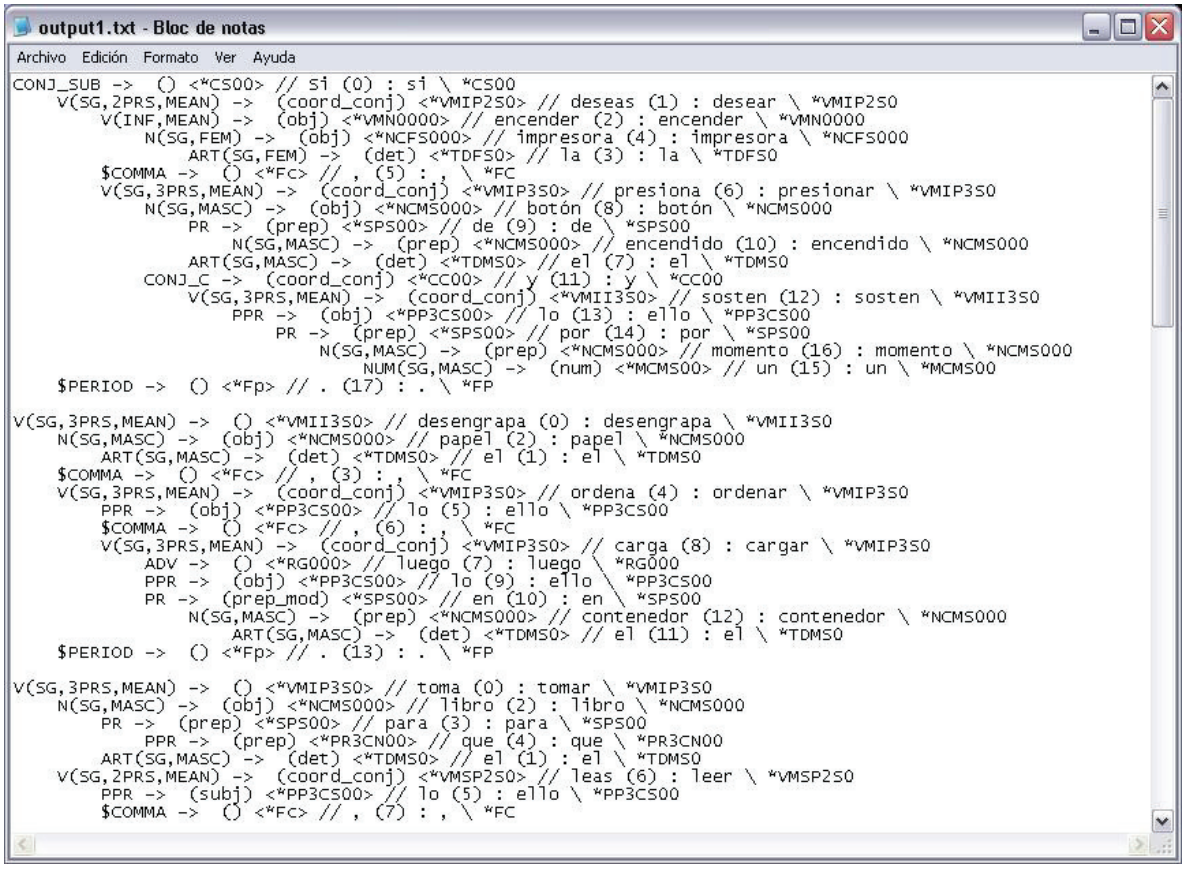
a través del cual se nos presentará un cuadro de diálogo que nos permitirá seleccionar el archivo de texto que será analizado, el cual deberá estar en formato de texto plano.



Una vez seleccionado el archivo de texto y presionado el botón “Abrir”, nuestro sistema utilizará el programa denominado «PARSER» que hará un análisis de la estructura sintáctica y morfológica de las oraciones contenidas en el archivo. Conforme el «PARSER» analiza el archivo, nos informará la cantidad de oraciones procesadas exitosamente.

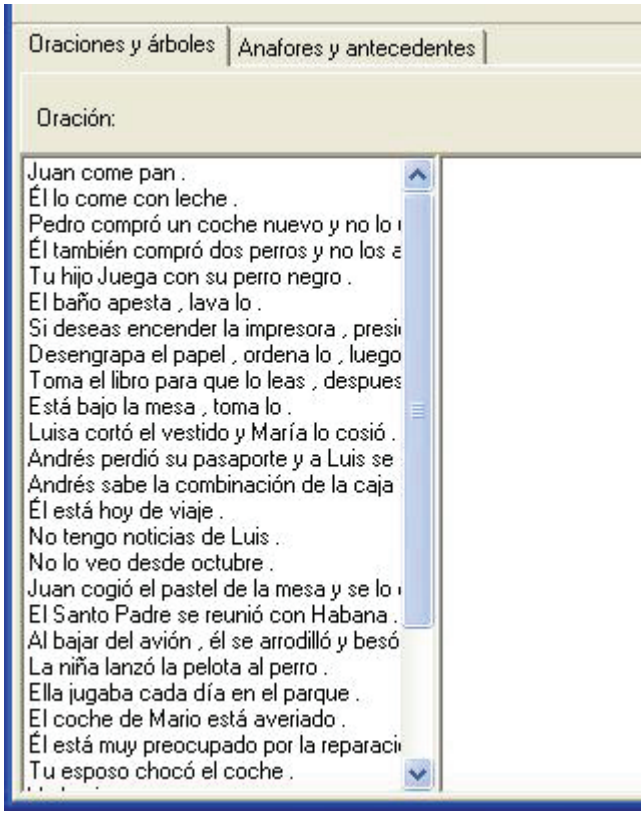
Una vez terminado el proceso de análisis y habiendo obtenido al menos una oración procesada exitosamente, se activará el botón “escribir árboles”.

Al presionar este botón, nuestro sistema grabará en un archivo de exto plano llamado “output1.txt”, los árboles de dependencias obtenidos del análisis, con la finalidad de corregir manualmente mediante algún editor de textos los errores que pudieran existir del análisis del parser.

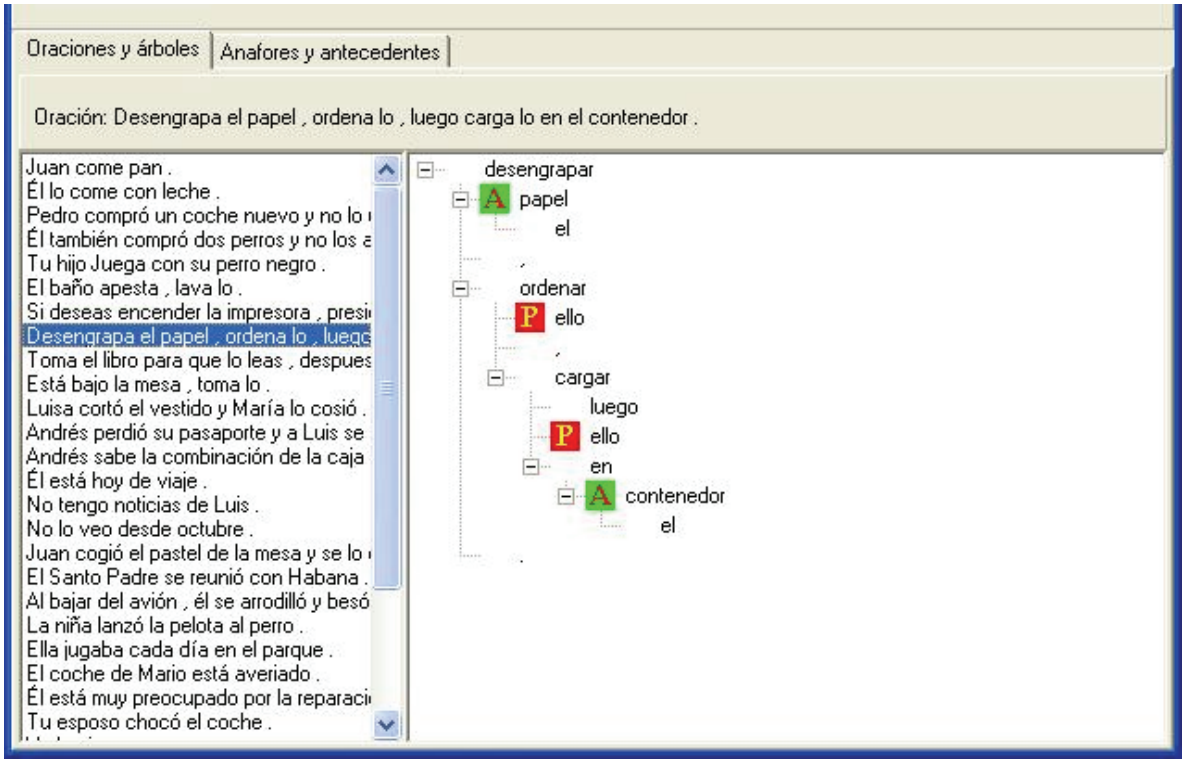


Una vez hechas y almacenadas las modificaciones necesarias a los árboles de dependencias, nuestro sistema podrá recuperarlos mediante el botón “Leer árboles”.

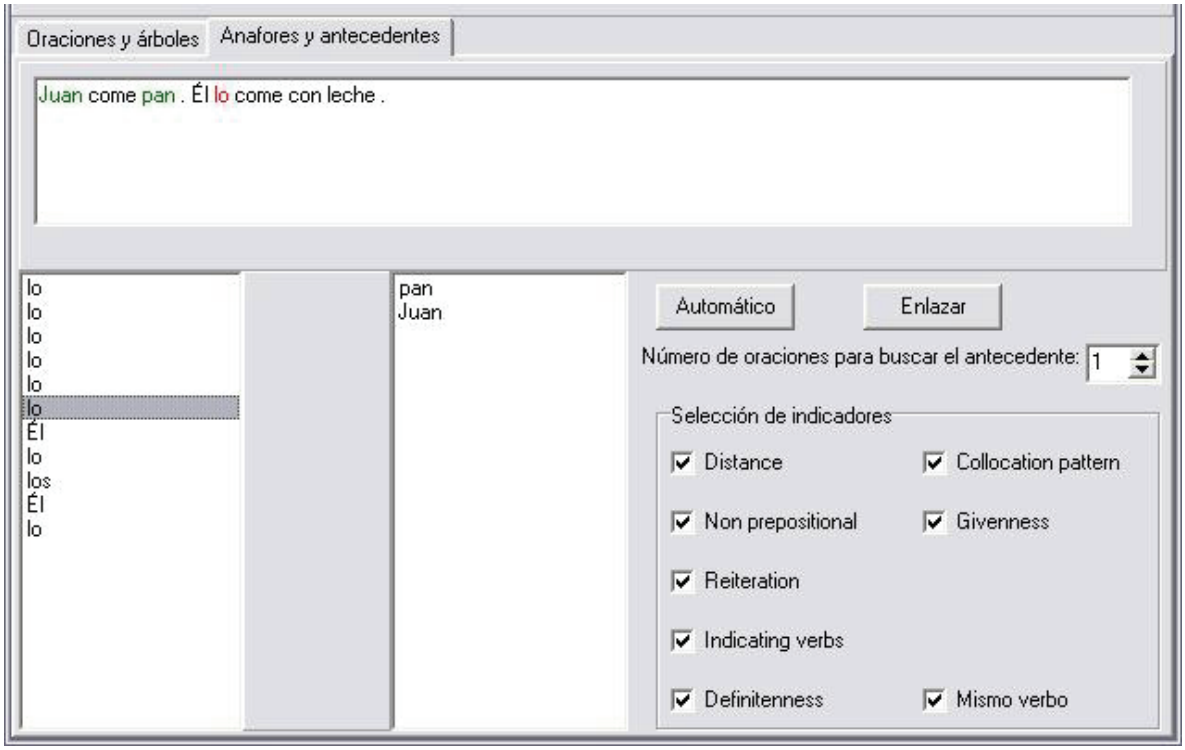
En unos de los paneles del sistema, aparecerán las oraciones recuperadas de los árboles de dependencias.



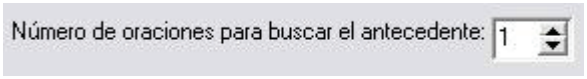
Nosotros podremos seleccionar cualquiera de esas oraciones y se nos mostrará el árbol de dependencias correspondiente. Este árbol nos indicará cuáles de sus elementos son pronombres (anafores) y cuales de ellos son frases nominales (candidatos a antecedentes).



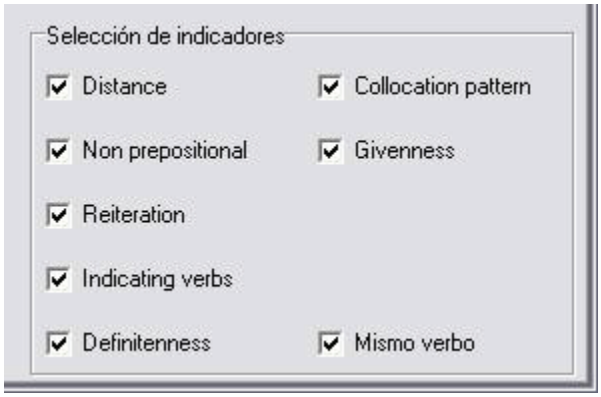
Si seleccionamos la pestaña “Anafores y antecedentes”, el sistema mostrará un panel que contiene 2 listas y una caja de textos. La lista del lado izquierdo nos muestra los anafores contenidos en todo el texto, la lista del lado derecho muestra los candidatos a antecedentes que corresponden al anafor seleccionado de la lista de anafores. En la caja de texto se nos muestran las oraciones en las que se encuentran esos elementos, el anafor en rojo y los candidatos en verde.



El número de oraciones en las que se buscarán los candidatos a antecedente se definirán en el siguiente control, que aceptará valores desde 0 hasta 10. Al seleccionar una cantidad diferente, se actualizará la lista de candidatos así como la caja de textos donde se muestran las oraciones que las contienen.

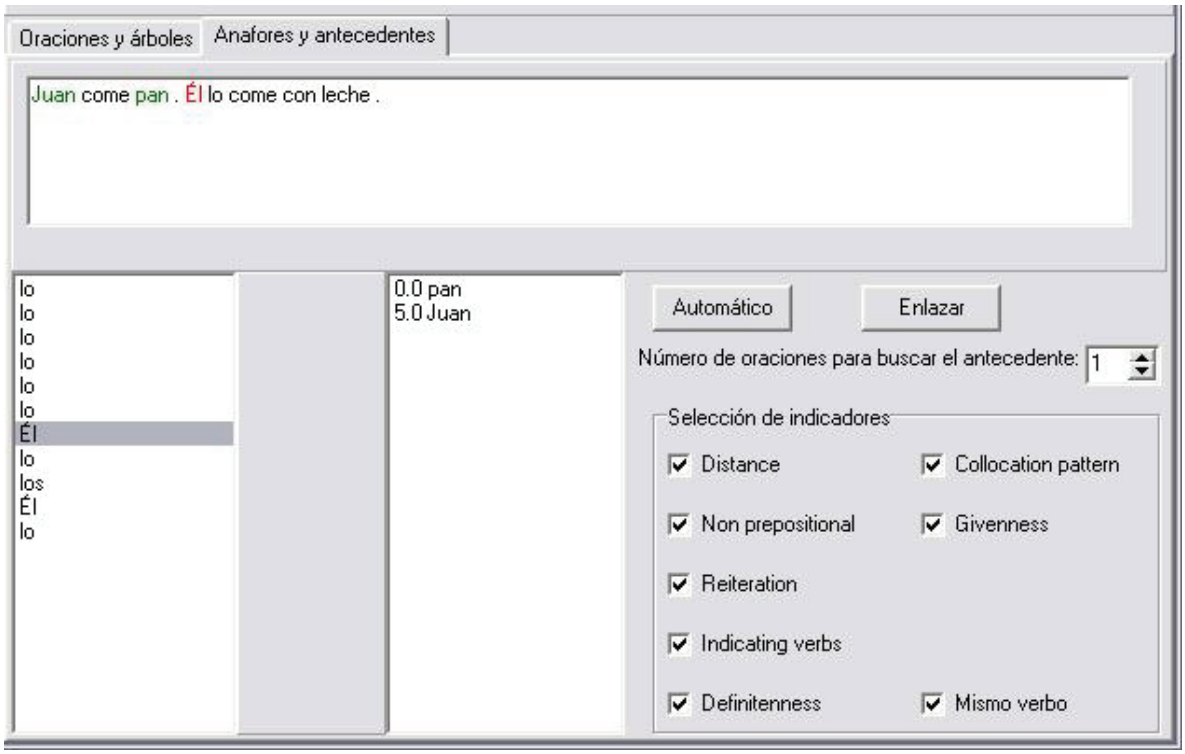


Una vez seleccionado el anafor a solucionar, el sistema da la posibilidad de seleccionar cuales son los indicadores que serán tomados en cuenta para la resolución de la anáfora.

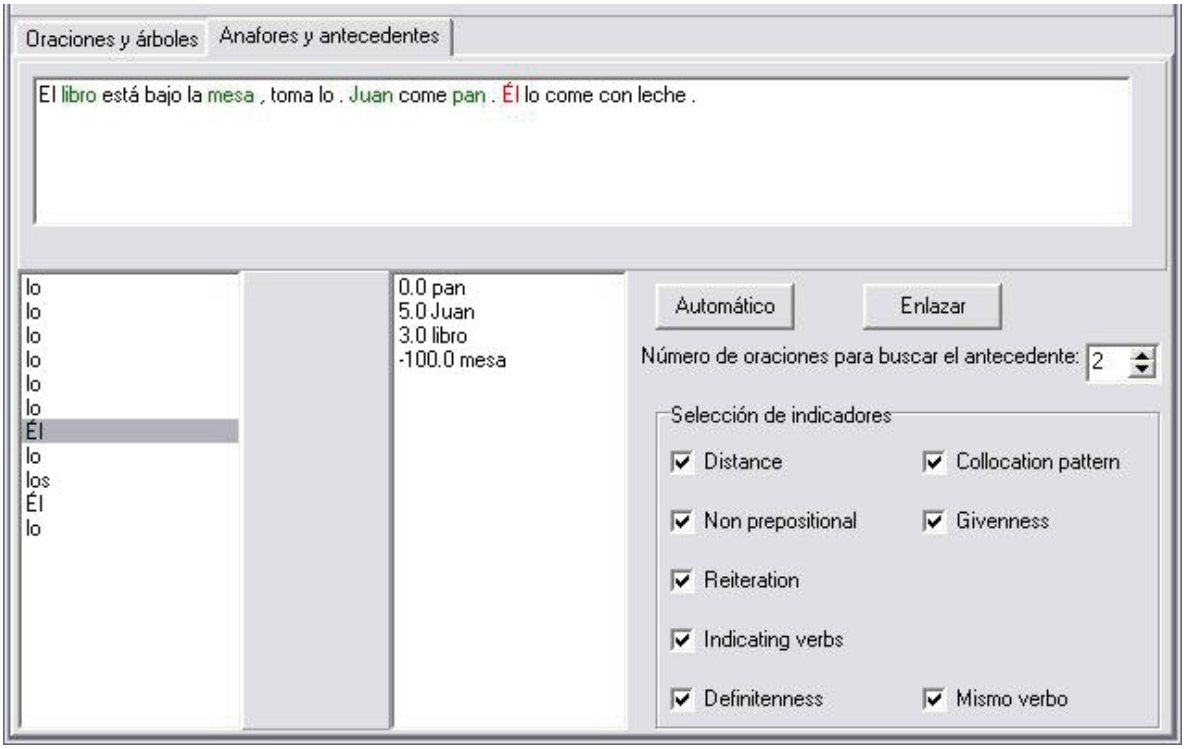


Una vez establecidos los indicadores de antecedencia, podemos hacer la resolución automática de la anáfora, presionando el botón “Automático”.

Al hacerlo, en la lista de candidatos, aparecerá la puntuación asignada a cada uno de los elementos.



En caso de que algún candidato no concuerde con el pronombre respecto a género y número, recibirá una puntuación de -100 y no se le aplicarán los factores de preferencia.



El candidato con la mejor puntuación es la solución ofrecida por el sistema.

CAPÍTULO 5 EXPERIMENTOS Y RESULTADOS

La evaluación de resultados se realizó de manera manual como se presenta en la siguiente sección.

5.1 Datos de prueba

Para los experimentos con el sistema, se tomó el texto que se muestra a continuación que consiste en 31 frases:

Juan come pan.
Él lo come con leche.
Pedro compró un coche nuevo y no lo usa.
Él también compró dos perros y no los alimenta.
Tu hijo Juega con su perro negro.
El baño apesta, lava lo.
Para encender la impresora, presiona el botón de encendido y sosten lo por un momento.
Desengrapa el papel, ordena lo, luego carga lo en el contenedor.
Toma el libro para que lo leas, después lo guardas.
Está bajo la mesa, toma lo.
Luisa cortó el vestido y María lo cosió.
Andrés perdió su pasaporte y a Luis se lo robaron.
Andrés sabe la combinación de la caja fuerte.
Él está hoy de viaje.
No tengo noticias de Luis.
No lo veo desde octubre.
Juan cogió el pastel de la mesa y se lo comió.
El Santo Padre se reunió con Fidel en La Habana.
Al bajar del avión, él se arrodilló y besó suelo cubano.
La niña lanzó la pelota al perro.
Ella jugaba cada día en el parque.
El coche de Mario está averiado.
Él está muy preocupado por la reparación.
Tu esposo chocó el coche.
Yo lo vi.
El chico de pelo rojo compró un libro.
Él lo compró en la tienda de Juan.
Luis compró varios libros.
Los compró en la tienda de Carlos.
Arturo compró un pantalón.
Él lo estrenó en una fiesta.

Posteriormente, fue procesado mediante el sistema y se obtuvo el siguiente resultado:

```
V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // come (1) : comer \ *VMIP3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // pan (2) : pan \ *NCMS000
    N(SG,FEM) ->  (subj) <*NP00000> // Juan (0) : juan \ *NP00000
    $PERIOD ->  () <*Fp> // . (3) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // come (2) : comer \ *VMIP3S0
    PR ->  (prep_mod) <*SPS00> // con (3) : con \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // leche (4) : leche \ *NCFS000
    PPR ->  () <*PP3CS00> // lo (1) : ello \ *PP3CS00
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

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,FEM) ->  (subj) <*NP00000> // Pedro (0) : pedro \ *NP00000
    CONJ_C ->  (coord_conj) <*CC00> // y (5) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // usa (8) : usar \
*VMIP3S0
            PPR ->  () <*PP3CS00> // lo (7) : ello \ *PP3CS00
            ADV ->  () <*RG000> // no (6) : no \ *RG000
    $PERIOD ->  () <*Fp> // . (9) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (2) : comprar \ *VMIS3S0
    ADV ->  () <*RG000> // también (1) : también \ *RG000
    CONJ_C ->  (obj) <*CC00> // y (5) : y \ *CC00
        N(PL,FEM) ->  () <*NCFP000> // perros (4) : perro \ *NCFP000
            NUM(PL,FEM) ->  (num) <*MCCP00> // dos (3) : dos \ *MCCP00
        N(SG,MASC) ->  (coord_conj) <*NCMS000> // no (6) : no \ *NCMS000
    PPR ->  () <*PP3MS00> // Él (0) : él \ *PP3MS00
    N(PL,MASC) ->  (obj) <*NCMP000> // los (7) : lo \ *NCMP000
    V(SG,3PRS,MEAN)  ->  (coord_conj)  <*VMIP3S0>  //  alimenta  (8)  :
alimentar \ *VMIP3S0
    $PERIOD ->  () <*Fp> // . (9) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Juega (2) : jugar \ *VMIP3S0
    PR ->  (prep_mod) <*SPS00> // con (3) : con \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // perro (5) : perro \ *NCFS000
            DET(SG,FEM) ->  (det) <*DP3CS00> // su (4) : su \ *DP3CS00
    N(SG,MASC) ->  (obj) <*NCMS000> // negro (6) : negro \ *NCMS000
    N(SG,MASC) ->  (subj) <*NCMS000> // hijo (1) : hijo \ *NCMS000
        DET(SG,MASC) ->  (det) <*DP2CS00> // Tu (0) : tu \ *DP2CS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // apesta (2) : apestar \ *VMIP3S0
    N(SG,MASC) ->  (subj) <*NCMS000> // baño (1) : baño \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    $COMMA ->  () <*Fc> // , (3) : , \ *FC
    V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // lava (4) : lavar \
*VMIP3S0
        PPR ->  (obj) <*PP3CS00> // lo (5) : ello \ *PP3CS00
```

```
    $PERIOD ->  () <*Fp> // . (6) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMSP3S0> // Para (0) : parir \ *VMSP3S0
    V(INF,MEAN)  ->    (obj)  <*VMN0000>  //  encender  (1)  :  encender  \
*VMN0000
        N(SG,FEM)  ->   (obj)  <*NCFS000>  //  impresora  (3)  :  impresora  \
*NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (2) : la \ *TDFS0
    $COMMA ->  () <*Fc> // , (4) : , \ *FC
    V(SG,3PRS,MEAN)  ->    (coord_conj)  <*VMIP3S0>  //  presiona  (5)  :
presionar \ *VMIP3S0
        N(SG,MASC) ->  (obj) <*NCMS000> // botón (7) : botón \ *NCMS000
            PR ->  (prep) <*SPS00> // de (8) : de \ *SPS00
                N(SG,MASC)  ->    (prep)  <*NCMS000>  //  encendido  (9)  :
encendido \ *NCMS000
            ART(SG,MASC) ->  (det) <*TDMS0> // el (6) : el \ *TDMS0
        CONJ_C ->  (coord_conj) <*CC00> // y (10) : y \ *CC00
            V(SG,3PRS,MEAN) ->  (coord_conj) <*VMII3S0> // sosten (11) :
sosten \ *VMII3S0
                PPR ->  (obj) <*PP3CS00> // lo (12) : ello \ *PP3CS00
                    PR ->  (prep) <*SPS00> // por (13) : por \ *SPS00
                        N(SG,MASC) ->  (prep) <*NCMS000> // momento (15)
: momento \ *NCMS000
                            NUM(SG,MASC) ->  (num) <*MCMS00> // un (14) :
un \ *MCMS00
    $PERIOD ->  () <*Fp> // . (16) : . \ *FP

V(SG,3PRS,MEAN)  ->    ()  <*VMII3S0>  //  Desengrapa  (0)  :  desengrapa  \
*VMII3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // papel (2) : papel \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // el (1) : el \ *TDMS0
    $COMMA ->  () <*Fc> // , (3) : , \ *FC
    V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // ordena (4) : ordenar \
*VMIP3S0
        PPR ->  (obj) <*PP3CS00> // lo (5) : ello \ *PP3CS00
        $COMMA ->  () <*Fc> // , (6) : , \ *FC
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // carga (8) : cargar
\ *VMIP3S0
            ADV ->  () <*RG000> // luego (7) : luego \ *RG000
            PPR ->  (obj) <*PP3CS00> // lo (9) : ello \ *PP3CS00
            PR ->  (prep_mod) <*SPS00> // en (10) : en \ *SPS00
                N(SG,MASC)  ->   (prep)  <*NCMS000>  //  contenedor  (12)  :
contenedor \ *NCMS000
                    ART(SG,MASC)  ->   (det)  <*TDMS0>  //  el  (11)  :  el  \
*TDMS0
    $PERIOD ->  () <*Fp> // . (13) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Toma (0) : tomar \ *VMIP3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // libro (2) : libro \ *NCMS000
        PR ->  (prep) <*SPS00> // para (3) : para \ *SPS00
            PPR ->  (prep) <*PR3CN00> // que (4) : que \ *PR3CN00
        ART(SG,MASC) ->  (det) <*TDMS0> // el (1) : el \ *TDMS0
    V(SG,2PRS,MEAN)  ->   (coord_conj)  <*VMSP2S0>  //  leas  (6)  :  leer  \
*VMSP2S0
        PPR ->  (subj) <*PP3CS00> // lo (5) : ello \ *PP3CS00
        $COMMA ->  () <*Fc> // , (7) : , \ *FC
```

```
        V(SG,3PRS,MEAN)  ->    (coord_conj)  <*VMII3S0>  //  despues  (8)  :
despues \ *VMII3S0
            PPR ->  (obj) <*PP3CS00> // lo (9) : ello \ *PP3CS00
            V(SG,2PRS,MEAN) ->  (coord_conj) <*VMIP2S0> // guardas (10) :
guardar \ *VMIP2S0
    $PERIOD ->  () <*Fp> // . (11) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Está (0) : estar \ *VMIP3S0
    PR ->  (cir) <*SPS00> // bajo (1) : bajo \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // mesa (3) : mesa \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (2) : la \ *TDFS0
    $COMMA ->  () <*Fc> // , (4) : , \ *FC
    V(SG,3PRS,MEAN)  ->    (coord_conj)  <*VMIP3S0>  //  toma  (5)  :  tomar  \
*VMIP3S0
        PPR ->  (obj) <*PP3CS00> // lo (6) : ello \ *PP3CS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // cortó (1) : cortar \ *VMIS3S0
    CONJ_C ->  (obj) <*CC00> // y (4) : y \ *CC00
        N(SG,MASC) ->  () <*NCMS000> // vestido (3) : vestido \ *NCMS000
            ART(SG,MASC) ->  (det) <*TDMS0> // el (2) : el \ *TDMS0
        N(PL,FEM)  ->    (coord_conj)  <*NP00000>  //  María  (5)  :  maría  \
*NP00000
    N(SG,FEM) ->  (subj) <*NP00000> // Luisa (0) : luisa \ *NP00000
    V(SG,3PRS,MEAN)  ->    (coord_conj)  <*VMIS3S0>  //  cosió  (7)  :  coser  \
*VMIS3S0
        PPR ->  () <*PP3CS00> // lo (6) : ello \ *PP3CS00
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // perdió (1) : perder \ *VMIS3S0
    N(SG,MASC)  ->    (obj)  <*NCMS000>  //  pasaporte  (3)  :  pasaporte  \
*NCMS000
        DET(SG,MASC) ->  (det) <*DP3CS00> // su (2) : su \ *DP3CS00
    N(SG,FEM) ->  (subj) <*NP00000> // Andrés (0) : andrés \ *NP00000
    CONJ_C ->  (coord_conj) <*CC00> // y (4) : y \ *CC00
        V(PL,3PRS,MEAN)  ->    (coord_conj)  <*VMIS3P0>  //  robaron  (9)  :
robar \ *VMIS3P0
            PPR ->  () <*PP3CS00> // lo (8) : ello \ *PP3CS00
            PPR ->  (subj) <*PP3CNR0> // se (7) : él \ *PP3CNR0
            PR ->  (cir) <*SPS00> // a (5) : a \ *SPS00
                N(PL,FEM)  ->    (prep)  <*NP00000>  //  Luis  (6)  :  luis  \
*NP00000
    $PERIOD ->  () <*Fp> // . (10) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // sabe (1) : saber \ *VMIP3S0
    N(SG,FEM)  ->    (obj)  <*NCFS000>  //  combinación  (3)  :  combinación  \
*NCFS000
        PR ->  (prep) <*SPS00> // de (4) : de \ *SPS00
            N(SG,FEM) ->  (prep) <*NCFS000> // caja (6) : caja \ *NCFS000
                ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
        ART(SG,FEM) ->  (det) <*TDFS0> // la (2) : la \ *TDFS0
    ADV ->  (cir) <*RG000> // fuerte (7) : fuerte \ *RG000
    N(SG,FEM) ->  (subj) <*NP00000> // Andrés (0) : andrés \ *NP00000
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // está (1) : estar \ *VMIP3S0
```

```
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    ADV ->  (cir) <*RG000> // hoy (2) : hoy \ *RG000
    PR ->  (cir) <*SPS00> // de (3) : de \ *SPS00
        N(SG,MASC) ->  (prep) <*NCMS000> // viaje (4) : viaje \ *NCMS000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMIP1S0> // tengo (1) : tener \ *VMIP1S0
    ADV ->  () <*RG000> // No (0) : no \ *RG000
    N(PL,FEM) ->  (obj) <*NCFP000> // noticias (2) : noticia \ *NCFP000
        PR ->  (prep) <*SPS00> // de (3) : de \ *SPS00
            N(PL,FEM) ->  (prep) <*NP00000> // Luis (4) : luis \ *NP00000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMIP1S0> // veo (2) : ver \ *VMIP1S0
    PPR ->  () <*PP3CS00> // lo (1) : ello \ *PP3CS00
    ADV ->  () <*RG000> // No (0) : no \ *RG000
    PR ->  (prep_mod) <*SPS00> // desde (3) : desde \ *SPS00
        N(SG,MASC) ->  (prep) <*NCMS000> // octubre (4) : octubre \
*NCMS000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // cogió (1) : coger \ *VMIS3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // pastel (3) : pastel \ *NCMS000
        PR ->  (prep) <*SPS00> // de (4) : de \ *SPS00
            N(SG,FEM) ->  (prep) <*NCFS000> // mesa (6) : mesa \ *NCFS000
                ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
        ART(SG,MASC) ->  (det) <*TDMS0> // el (2) : el \ *TDMS0
    N(SG,FEM) ->  (subj) <*NP00000> // Juan (0) : juan \ *NP00000
    CONJ_C ->  (coord_conj) <*CC00> // y (7) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIS3S0> // comió (10) : comer
\ *VMIS3S0
            PPR ->  () <*PP3CNR0> // se (8) : él \ *PP3CNR0
            PPR ->  () <*PP3CS00> // lo (9) : ello \ *PP3CS00
    $PERIOD ->  () <*Fp> // . (11) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // arrodilló (7) : arrodillar \
*VMIS3S0
    PPR ->  () <*PP3CNR0> // se (6) : él \ *PP3CNR0
    PPR ->  (subj) <*PP3MS00> // él (5) : él \ *PP3MS00
    PR ->  (cir) <*SPCMS> // del (2) : del \ *SPCMS
        N(SG,MASC) ->  (prep) <*NCMS000> // avión (3) : avión \ *NCMS000
    $COMMA ->  (cir) <*Fc> // , (4) : , \ *FC
    PR ->  (cir) <*SPCMS> // Al (0) : al \ *SPCMS
        V(INF,MEAN) ->  (prep) <*VMN0000> // bajar (1) : bajar \ *VMN0000
    CONJ_C ->  (coord_conj) <*CC00> // y (8) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIS3S0> // besó (9) : besar \
*VMIS3S0
            V(SG,1PRS,MEAN) ->  (coord_conj) <*VMIP1S0> // suelo (10) :
soler \ *VMIP1S0
                N(SG,MASC) ->  (obj) <*NCMS000> // cubano (11) : cubano \
*NCMS000
    $PERIOD ->  () <*Fp> // . (12) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // lanzó (2) : lanzar \ *VMIS3S0
    N(SG,FEM) ->  (obj) <*NCFS000> // pelota (4) : pelota \ *NCFS000
        PR ->  (prep) <*SPCMS> // al (5) : al \ *SPCMS
```

```
            N(SG,FEM)  ->    (prep)  <*NCFS000>  //  perro  (6)  :  perro  \
*NCFS000
        ART(SG,FEM) ->  (det) <*TDFS0> // la (3) : la \ *TDFS0
    N(SG,FEM) ->  (subj) <*NCFS000> // niña (1) : niña \ *NCFS000
        ART(SG,FEM) ->  (det) <*TDFS0> // La (0) : la \ *TDFS0
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMII3S0> // jugaba (1) : jugar \ *VMII3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // día (3) : día \ *NCMS000
        PR ->  (prep) <*SPS00> // en (4) : en \ *SPS00
            N(SG,MASC)  ->    (prep)  <*NCMS000>  //  parque  (6)  :  parque  \
*NCMS000
                ART(SG,MASC) ->  (det) <*TDMS0> // el (5) : el \ *TDMS0
        DET(SG,MASC) ->  (det) <*DI0CS00> // cada (2) : cada \ *DI0CS00
    PPR ->  () <*PP3FS00> // Ella (0) : ella \ *PP3FS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

#*$$estar(SG,1PRS)# ->  () <*$$estar> // está (4) : estar \ *VMIP3S0
    PART(SG,MASC) ->  () <*VMPP0SM> // averiado (5) : averiar \ *VMPP0SM
    N(SG,MASC) ->  (subj) <*NCMS000> // coche (1) : coche \ *NCMS000
        PR ->  (prep) <*SPS00> // de (2) : de \ *SPS00
            N(PL,FEM)  ->    (prep)  <*NP00000>  //  Mario  (3)  :  mario  \
*NP00000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    $PERIOD ->  () <*Fp> // . (6) : . \ *FP

#*$$estar(SG,1PRS)# ->  () <*$$estar> // está (1) : estar \ *VMIP3S0
    ADV ->  (adver) <*RG000> // muy (2) : muy \ *RG000
    PART(SG,MASC) ->  (adver) <*VMPP0SM> // preocupado (3) : preocupar \
*VMPP0SM
    PR ->  (prep_mod) <*SPS00> // por (4) : por \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // reparación (6) : reparación \
*NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // chocó (2) : chocar \ *VMIS3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // coche (4) : coche \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // el (3) : el \ *TDMS0
    N(SG,MASC) ->  (subj) <*NCMS000> // esposo (1) : esposo \ *NCMS000
        DET(SG,MASC) ->  (det) <*DP2CS00> // Tu (0) : tu \ *DP2CS00
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMID1S0> // vi (2) : ver \ *VMID1S0
    PPR ->  () <*PP3CS00> // lo (1) : ello \ *PP3CS00
    PPR ->  () <*PP1CS00> // Yo (0) : yo \ *PP1CS00
    $PERIOD ->  () <*Fp> // . (3) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (5) : comprar \ *VMIS3S0
    NUM(SG,MASC) ->  (obj) <*MCMS00> // un (6) : un \ *MCMS00
    N(SG,MASC) ->  (subj) <*NCMS000> // chico (1) : chico \ *NCMS000
        PR ->  (prep) <*SPS00> // de (2) : de \ *SPS00
            N(SG,MASC)  ->    (prep)  <*NCMS000>  //  pelo  (3)  :  pelo  \
*NCMS000
```

```
                N(SG,MASC) ->   (comp) <*NCMS000> // rojo (4) : rojo \
*NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    V(SG,1PRS,MEAN) ->  (coord_conj) <*VMIP1S0> // libro (7) : librar \
*VMIP1S0
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (2) : comprar \ *VMIS3S0
    PPR ->  () <*PP3CS00> // lo (1) : ello \ *PP3CS00
    PR ->  (prep_mod) <*SPS00> // en (3) : en \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // tienda (5) : tienda \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (4) : la \ *TDFS0
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    PR ->  (cir) <*SPS00> // de (6) : de \ *SPS00
        N(PL,FEM) ->  (prep) <*NP00000> // Juan (7) : juan \ *NP00000
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    N(PL,MASC) ->  (obj) <*NCMP000> // libros (3) : libro \ *NCMP000
        DET(PL,MASC) ->   (det) <*DI3MP00> // varios (2) : varios \
*DI3MP00
    N(SG,FEM) ->  (subj) <*NP00000> // Luis (0) : luis \ *NP00000
    $PERIOD ->  () <*Fp> // . (4) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    PR ->  (prep_mod) <*SPS00> // en (2) : en \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // tienda (4) : tienda \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (3) : la \ *TDFS0
    PR ->  (prep_mod) <*SPS00> // de (5) : de \ *SPS00
        N(PL,FEM) ->  (prep) <*NP00000> // Carlos (6) : carlos \ *NP00000
    N(SG,MASC) ->  (subj) <*NCMS000> // Los (0) : los \ *NCMS000
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    NUM(SG,MASC) ->  (obj) <*MCMS00> // un (2) : un \ *MCMS00
        N(SG,FEM) ->  (poss_mod) <*NCFS000> // pantalón (3) : pantalón \
*NCFS000
    N(SG,FEM) ->  (subj) <*NP00000> // Arturo (0) : arturo \ *NP00000
    $PERIOD ->  () <*Fp> // . (4) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // estrenó (2) : estrenar \ *VMIS3S0
    PPR ->  () <*PP3CS00> // lo (1) : ello \ *PP3CS00
    PR ->  (prep_mod) <*SPS00> // en (3) : en \ *SPS00
        NUM(SG,FEM) ->  (prep) <*MCFS00> // una (4) : un \ *MCFS00
    N(SG,FEM) ->  (obj) <*NCFS000> // fiesta (5) : fiesta \ *NCFS000
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (6) : . \ *FP
```

Este resultado fue verificado manualmente, encontrándose y corrigiendo una serie de errores arrojados por el parser, a continuación se muestra el archivo corregido:

```
V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // come (1) : comer \ *VMIP3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // pan (2) : pan \ *NCMS000
    N(SG,MASC) ->  (subj) <*NPMS000> // Juan (0) : juan \ *NPMS000
    $PERIOD ->  () <*Fp> // . (3) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // come (2) : comer \ *VMIP3S0
    PR ->  (prep_mod) <*SPS00> // con (3) : con \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // leche (4) : leche \ *NCFS000
    PPR ->  (obj) <*PP3MS00> // lo (1) : ello \ *PP3MS00
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

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

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (2) : comprar \ *VMIS3S0
    ADV ->  () <*RG000> // también (1) : también \ *RG000
    CONJ_C ->  (obj) <*CC00> // y (5) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // alimenta (8) :
alimentar \ *VMIP3S0
            ADV ->  () <*RG000> // no (6) : no \ *NCMS000
            PPR ->  (obj) <*PP3MP00> // los (7) : ello \ *PP3MP00
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    N(PL,MASC) ->  (obj) <*NCMP000> // perros (4) : perro \ *NCMP000
        NUM(PL,FEM) ->  (num) <*MCCP00> // dos (3) : dos \ *MCCP00
    $PERIOD ->  () <*Fp> // . (9) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Juega (2) : jugar \ *VMIP3S0
    PR ->  (prep_mod) <*SPS00> // con (3) : con \ *SPS00
    N(SG,MASC) ->  (prep) <*NCMS000> // perro (5) : perro \ *NCMS000
        N(SG,MASC) ->  (obj) <*NCMS000> // negro (6) : negro \ *NCMS000
        DET(SG,FEM) ->  (det) <*DP3CS00> // su (4) : su \ *DP3CS00
    N(SG,MASC) ->  (subj) <*NCMS000> // hijo (1) : hijo \ *NCMS000
        DET(SG,MASC) ->  (det) <*DP2CS00> // Tu (0) : tu \ *DP2CS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // apesta (2) : apestar \ *VMIP3S0
    N(SG,MASC) ->  (subj) <*NCMS000> // baño (1) : baño \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    $COMMA ->  () <*Fc> // , (3) : , \ *FC
    V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // lava (4) : lavar \
*VMIP3S0
        PPR ->  (obj) <*PP3MS00> // lo (5) : ello \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (6) : . \ *FP

CONJ_SUB ->  () <*CS00> // Si (0) : si \ *CS00
```

```
    V(SG,2PRS,MEAN) ->  (coord_conj) <*VMIP2S0> // deseas (1) : desear \
*VMIP2S0
        V(INF,MEAN) ->   (obj) <*VMN0000> // encender (2) : encender \
*VMN0000
            N(SG,FEM) ->  (obj) <*NCFS000> // impresora (4) : impresora \
*NCFS000
                ART(SG,FEM) ->  (det) <*TDFS0> // la (3) : la \ *TDFS0
        $COMMA ->  () <*Fc> // , (5) : , \ *FC
        V(SG,3PRS,MEAN) ->   (coord_conj) <*VMIP3S0> // presiona (6) :
presionar \ *VMIP3S0
            N(SG,MASC) ->   (obj) <*NCMS000> // botón (8) : botón \
*NCMS000
                PR ->  (prep) <*SPS00> // de (9) : de \ *SPS00
                    N(SG,MASC) ->  (prep) <*NCMS000> // encendido (10) :
encendido \ *NCMS000
                ART(SG,MASC) ->  (det) <*TDMS0> // el (7) : el \ *TDMS0
            CONJ_C ->  (coord_conj) <*CC00> // y (11) : y \ *CC00
                V(SG,3PRS,MEAN) ->   (coord_conj) <*VMII3S0> // sosten
(12) : sosten \ *VMII3S0
                    PPR ->  (obj) <*PP3MS00> // lo (13) : ello \ *PP3MS00
                        PR ->  (prep) <*SPS00> // por (14) : por \ *SPS00
                            N(SG,MASC) ->   (prep) <*NCMS000> // momento
(16) : momento \ *NCMS000
                                NUM(SG,MASC) ->   (num) <*MCMS00> // un
(15) : un \ *MCMS00
    $PERIOD ->  () <*Fp> // . (17) : . \ *FP

V(SG,3PRS,MEAN) ->   () <*VMII3S0> // Desengrapa (0) : desengrapar \
*VMII3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // papel (2) : papel \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // el (1) : el \ *TDMS0
    $COMMA ->  () <*Fc> // , (3) : , \ *FC
    V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // ordena (4) : ordenar \
*VMIP3S0
        PPR ->  (obj) <*PP3MS00> // lo (5) : ello \ *PP3MS00
        $COMMA ->  () <*Fc> // , (6) : , \ *FC
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIP3S0> // carga (8) : cargar
\ *VMIP3S0
            ADV ->  () <*RG000> // luego (7) : luego \ *RG000
            PPR ->  (obj) <*PP3MS00> // lo (9) : ello \ *PP3MS00
            PR ->  (prep_mod) <*SPS00> // en (10) : en \ *SPS00
                N(SG,MASC) ->   (prep) <*NCMS000> // contenedor (12) :
contenedor \ *NCMS000
                    ART(SG,MASC) ->   (det) <*TDMS0> // el (11) : el \
*TDMS0
    $PERIOD ->  () <*Fp> // . (13) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Toma (0) : tomar \ *VMIP3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // libro (2) : libro \ *NCMS000
        PR ->  (prep) <*SPS00> // para (3) : para \ *SPS00
            PPR ->  (prep) <*PR3CN00> // que (4) : que \ *PR3CN00
        ART(SG,MASC) ->  (det) <*TDMS0> // el (1) : el \ *TDMS0
    V(SG,2PRS,MEAN) ->   (coord_conj) <*VMSP2S0> // leas (6) : leer \
*VMSP2S0
        PPR ->  (subj) <*PP3MS00> // lo (5) : ello \ *PP3MS00
        $COMMA ->  () <*Fc> // , (7) : , \ *FC
```

```
        V(SG,3PRS,MEAN)  ->    (coord_conj)  <*VMII3S0>  //  despues  (8)  :
despues \ *VMII3S0
            PPR ->  (obj) <*PP3MS00> // lo (9) : ello \ *PP3MS00
            V(SG,2PRS,MEAN) ->  (coord_conj) <*VMIP2S0> // guardas (10) :
guardar \ *VMIP2S0
    $PERIOD ->  () <*Fp> // . (11) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // Está (0) : estar \ *VMIP3S0
    PR ->  (cir) <*SPS00> // bajo (1) : bajo \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // mesa (3) : mesa \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (2) : la \ *TDFS0
    $COMMA ->  () <*Fc> // , (4) : , \ *FC
    V(SG,3PRS,MEAN)  ->   (coord_conj)  <*VMIP3S0>  //  toma  (5)  :  tomar  \
*VMIP3S0
        PPR ->  (obj) <*PP3MS00> // lo (6) : ello \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // cortó (1) : cortar \ *VMIS3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // vestido (3) : vestido \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // el (2) : el \ *TDMS0
    CONJ_C ->  (coord_conj) <*CC00> // y (4) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIS3S0> // cosió (7) : coser
\ *VMIS3S0
            PPR ->  (obj) <*PP3MS00> // lo (6) : ello \ *PP3MS00
        N(PL,FEM)  ->   (coord_conj)  <*NPFS000>  //  María  (5)  :  maría  \
*NP00000
    N(SG,FEM) ->  (subj) <*NPFS000> // Luisa (0) : luisa \ *NP00000
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // perdió (1) : perder \ *VMIS3S0
    N(SG,MASC)  ->   (obj)  <*NCMS000>  //  pasaporte  (3)  :  pasaporte  \
*NCMS000
        DET(SG,MASC) ->  (det) <*DP3CS00> // su (2) : su \ *DP3CS00
    N(SG,MASC) ->  (subj) <*NPMS000> // Andrés (0) : andrés \ *NPMS000
    CONJ_C ->  (coord_conj) <*CC00> // y (4) : y \ *CC00
        V(PL,3PRS,MEAN)  ->   (coord_conj)  <*VMIS3P0>  //  robaron  (9)  :
robar \ *VMIS3P0
            PPR ->  (obj) <*PP3MS00> // lo (8) : ello \ *PP3MS00
            PPR ->  (subj) <*PP3CNR0> // se (7) : él \ *PP3CNR0
            PR ->  (cir) <*SPS00> // a (5) : a \ *SPS00
                N(PL,MASC)  ->   (prep)  <*NPMS000>  //  Luis  (6)  :  luis  \
*NPMS000
    $PERIOD ->  () <*Fp> // . (10) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // sabe (1) : saber \ *VMIP3S0
    N(SG,FEM)  ->   (obj)  <*NCFS000>  //  combinación  (3)  :  combinación  \
*NCFS000
        PR ->  (prep) <*SPS00> // de (4) : de \ *SPS00
            N(SG,FEM) ->  (prep) <*NCFS000> // caja (6) : caja \ *NCFS000
                ADV ->  (cir) <*RG000> // fuerte (7) : fuerte \ *RG000
                ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
        ART(SG,FEM) ->  (det) <*TDFS0> // la (2) : la \ *TDFS0
    N(SG,MASC) ->  (subj) <*NPMS000> // Andrés (0) : andrés \ *NPMS000
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIP3S0> // está (1) : estar \ *VMIP3S0
```

```
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    ADV ->  (cir) <*RG000> // hoy (2) : hoy \ *RG000
    PR ->  (cir) <*SPS00> // de (3) : de \ *SPS00
        N(SG,MASC) ->  (prep) <*NCMS000> // viaje (4) : viaje \ *NCMS000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMIP1S0> // tengo (1) : tener \ *VMIP1S0
    ADV ->  () <*RG000> // No (0) : no \ *RG000
    N(PL,FEM) ->  (obj) <*NCFP000> // noticias (2) : noticia \ *NCFP000
        PR ->  (prep) <*SPS00> // de (3) : de \ *SPS00
            N(PL,MASC) ->  (prep) <*NPMS000> // Luis (4) : luis \
*NPMS000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMIP1S0> // veo (2) : ver \ *VMIP1S0
    PPR ->  (obj) <*PP3MS00> // lo (1) : ello \ *PP3MS00
    ADV ->  () <*RG000> // No (0) : no \ *RG000
    PR ->  (prep_mod) <*SPS00> // desde (3) : desde \ *SPS00
        N(SG,MASC) ->  (prep) <*NCMS000> // octubre (4) : octubre \
*NCMS000
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // cogió (1) : coger \ *VMIS3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // pastel (3) : pastel \ *NCMS000
    PR ->  (prep) <*SPS00> // de (4) : de \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // mesa (6) : mesa \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
        ART(SG,MASC) ->  (det) <*TDMS0> // el (2) : el \ *TDMS0
    N(SG,FEM) ->  (subj) <*NPMS000> // Juan (0) : juan \ *NPMS000
    CONJ_C ->  (coord_conj) <*CC00> // y (7) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIS3S0> // comió (10) : comer
\ *VMIS3S0
            PPR ->  (subj) <*PP3CNR0> // se (8) : él \ *PP3CNR0
            PPR ->  (obj) <*PP3MS00> // lo (9) : ello \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (11) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // reunió (4) : reunir \ *VMIS3S0
    PPR ->  () <*PP3CNR0> // se (3) : él \ *PP3CNR0
    N(SG,MASC) ->  (subj) <*NCMS000> // Padre (2) : padre \ *NCMS000
        N(SG,MASC) ->  (comp) <*NCMS000> // Santo (1) : santo \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    PR ->  (cir) <*SPS00> // con (5) : con \ *SPS00
        N(PL,MASC) ->  (prep) <*NPMS000> // Fidel (6) : fidel \ *NPMS000
    PR ->  (prep) <*SPS00> // en (4) : en \ *SPS00
        N(SG,FEM) ->  (prep) <*NPFS000> // Habana (6) : Habana \ *NPFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // La (5) : la \ *TDFS0
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // arrodilló (7) : arrodillar \
*VMIS3S0
    PPR ->  () <*PP3CNR0> // se (6) : él \ *PP3CNR0
    PPR ->  (subj) <*PP3MS00> // él (5) : él \ *PP3MS00
    PR ->  (cir) <*SPCMS> // del (2) : del \ *SPCMS
        N(SG,MASC) ->  (prep) <*NCMS000> // avión (3) : avión \ *NCMS000
    $COMMA ->  (cir) <*Fc> // , (4) : , \ *FC
    PR ->  (cir) <*SPCMS> // Al (0) : al \ *SPCMS
```

```
        V(INF,MEAN) ->  (prep) <*VMN0000> // bajar (1) : bajar \ *VMN0000
    CONJ_C ->  (coord_conj) <*CC00> // y (8) : y \ *CC00
        V(SG,3PRS,MEAN) ->  (coord_conj) <*VMIS3S0> // besó (9) : besar \
*VMIS3S0
            N(SG,MASC) ->   (obj) <*NCMS000> // suelo (10) : suelo \
*NCMS000
                N(SG,MASC) ->  (comp) <*NCMS000> // cubano (11) : cubano
\ *NCMS000
    $PERIOD ->  () <*Fp> // . (12) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // lanzó (2) : lanzar \ *VMIS3S0
    N(SG,FEM) ->  (obj) <*NCFS000> // pelota (4) : pelota \ *NCFS000
        PR ->  (prep) <*SPCMS> // al (5) : al \ *SPCMS
            N(SG,FEM) ->   (prep) <*NCMS000> // perro (6) : perro \
*NCFS000
        ART(SG,FEM) ->  (det) <*TDFS0> // la (3) : la \ *TDFS0
    N(SG,FEM) ->  (subj) <*NCFS000> // niña (1) : niña \ *NCFS000
        ART(SG,FEM) ->  (det) <*TDFS0> // La (0) : la \ *TDFS0
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMII3S0> // jugaba (1) : jugar \ *VMII3S0
    PR ->  (prep) <*SPS00> // en (4) : en \ *SPS00
        N(SG,MASC) ->   (prep) <*NCMS000> // parque (6) : parque \
*NCMS000
            ART(SG,MASC) ->  (det) <*TDMS0> // el (5) : el \ *TDMS0
    DET(SG,MASC) ->  (det) <*DI0CS00> // cada (2) : cada \ *DI0CS00
        N(SG,MASC) ->  () <*NCMS000> // día (3) : día \ *NCMS000
    PPR ->  (subj) <*PP3FS00> // Ella (0) : ella \ *PP3FS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

#*$$estar(SG,1PRS)# ->  () <*$$estar> // está (4) : estar \ *VMIP3S0
    PART(SG,MASC) ->  () <*VMPP0SM> // averiado (5) : averiar \ *VMPP0SM
    N(SG,MASC) ->  (subj) <*NCMS000> // coche (1) : coche \ *NCMS000
        PR ->  (prep) <*SPS00> // de (2) : de \ *SPS00
            N(PL,MASC) ->   (prep) <*NPMS000> // Mario (3) : mario \
*NPMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    $PERIOD ->  () <*Fp> // . (6) : . \ *FP

#*$$estar(SG,1PRS)# ->  () <*$$estar> // está (1) : estar \ *VMIP3S0
    ADV ->  (adver) <*RG000> // muy (2) : muy \ *RG000
    PART(SG,MASC) ->  (adver) <*VMPP0SM> // preocupado (3) : preocupar \
*VMPP0SM
    PR ->  (prep_mod) <*SPS00> // por (4) : por \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // reparación (6) : reparación \
*NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (5) : la \ *TDFS0
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,1PRS,MEAN) ->  () <*VMIP1S0> // chocó (2) : chocar \ *VMIP1S0
    N(SG,MASC) ->  (obj) <*NCMS000> // coche (4) : coche \ *NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // el (3) : el \ *TDMS0
    N(SG,MASC) ->  (subj) <*NCMS000> // esposo (1) : esposo \ *NCMS000
        DET(SG,MASC) ->  (det) <*DP2CS00> // Tu (0) : tu \ *DP2CS00
    $PERIOD ->  () <*Fp> // . (5) : . \ *FP
```

```
V(SG,1PRS,MEAN) ->  () <*VMID1S0> // vi (2) : ver \ *VMID1S0
    PPR ->  (obj) <*PP3MS00> // lo (1) : ello \ *PP3MS00
    PPR ->  (subj) <*PP1CS00> // Yo (0) : yo \ *PP1CS00
    $PERIOD ->  () <*Fp> // . (3) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (5) : comprar \ *VMIS3S0
    N(SG,MASC) ->  (obj) <*NCMS000> // libro (7) : libro \ *NCMS000
        NUM(SG,MASC) ->  (obj) <*MCMS00> // un (6) : un \ *MCMS00
    N(SG,MASC) ->  (subj) <*NCMS000> // chico (1) : chico \ *NCMS000
        PR ->  (prep) <*SPS00> // de (2) : de \ *SPS00
            N(SG,MASC) ->  (prep) <*NCMS000> // pelo (3) : pelo \
*NCMS000
                N(SG,MASC) ->  (comp) <*NCMS000> // rojo (4) : rojo \
*NCMS000
        ART(SG,MASC) ->  (det) <*TDMS0> // El (0) : el \ *TDMS0
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (2) : comprar \ *VMIS3S0
    PPR ->  (obj) <*PP3MS00> // lo (1) : ello \ *PP3MS00
    PR ->  (prep_mod) <*SPS00> // en (3) : en \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // tienda (5) : tienda \ *NCFS000
            PR ->  (cir) <*SPS00> // de (6) : de \ *SPS00
                N(PL,MASC) ->  (prep) <*NPMS000> // Juan (7) : juan \
*NPMS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (4) : la \ *TDFS0
    PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
    $PERIOD ->  () <*Fp> // . (8) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    N(PL,MASC) ->  (obj) <*NCMP000> // libros (3) : libro \ *NCMP000
        DET(PL,MASC) ->  (det) <*DI3MP00> // varios (2) : varios \
*DI3MP00
    N(SG,MASC) ->  (subj) <*NPMS000> // Luis (0) : luis \ *NPMS000
    $PERIOD ->  () <*Fp> // . (4) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    PR ->  (prep_mod) <*SPS00> // en (2) : en \ *SPS00
        N(SG,FEM) ->  (prep) <*NCFS000> // tienda (4) : tienda \ *NCFS000
            ART(SG,FEM) ->  (det) <*TDFS0> // la (3) : la \ *TDFS0
            PR ->  (prep_mod) <*SPS00> // de (5) : de \ *SPS00
                N(PL,MASC) ->  (prep) <*NPMS000> // Carlos (6) : carlos \
*NPMS000
    PPR ->  (subj) <*PP3MP00> // Los (0) : ello \ *PP3MP00
    $PERIOD ->  () <*Fp> // . (7) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // compró (1) : comprar \ *VMIS3S0
    N(SG,MASC) ->  (poss_mod) <*NCMS000> // pantalón (3) : pantalón \
*NCMS000
        NUM(SG,MASC) ->  (obj) <*MCMS00> // un (2) : un \ *MCMS00
    N(SG,MASC) ->  (subj) <*NPMS000> // Arturo (0) : arturo \ *NPMS000
    $PERIOD ->  () <*Fp> // . (4) : . \ *FP

V(SG,3PRS,MEAN) ->  () <*VMIS3S0> // estrenó (2) : estrenar \ *VMIS3S0
    PPR ->  (obj) <*PP3MS00> // lo (1) : ello \ *PP3MS00
    PR ->  (prep_mod) <*SPS00> // en (3) : en \ *SPS00
```

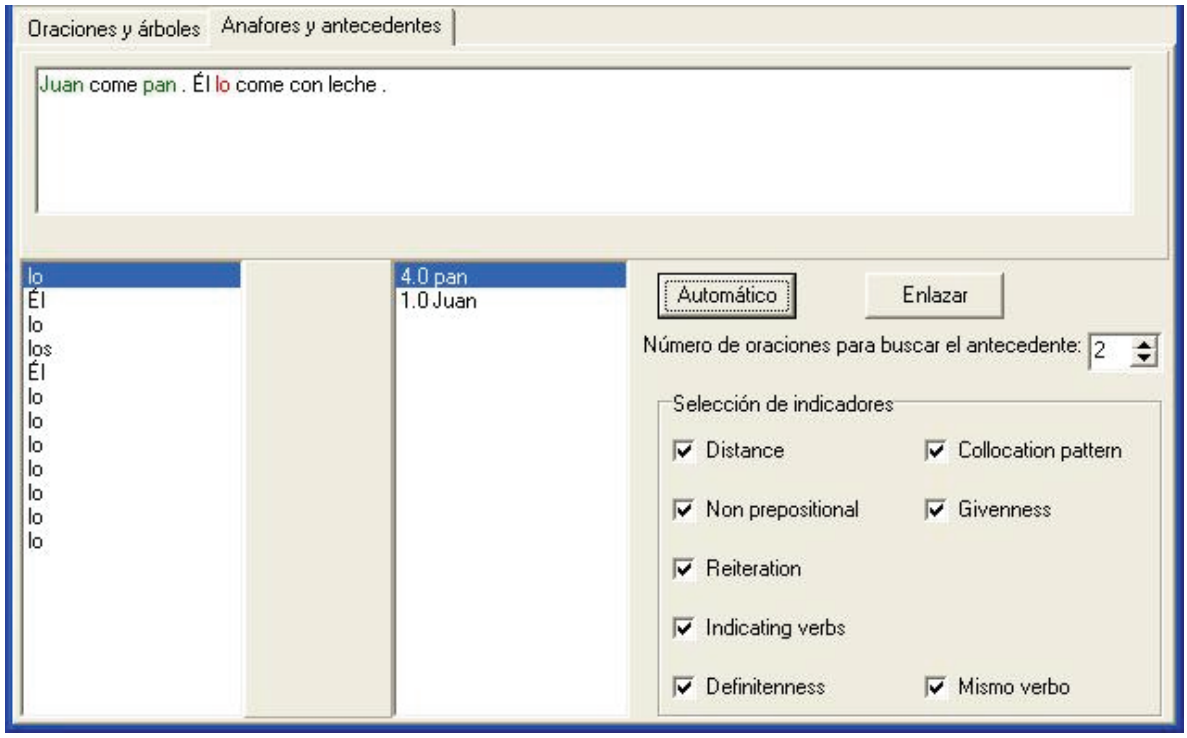
```
    N(SG,FEM) ->  (obj) <*NCFS000> // fiesta (5) : fiesta \ *NCFS000
        NUM(SG,FEM) ->  (prep) <*MCFS00> // una (4) : un \ *MCFS00
PPR ->  (subj) <*PP3MS00> // Él (0) : él \ *PP3MS00
$PERIOD ->  () <*Fp> // . (6) : . \ *FP
```

Una vez almacenado el archivo modificado, se le indicó al sistema que recupera la información de éste, encontrando en total 25 anafores con sus respectivos candidatos a antecedente.

5.2 Resultados experimentales

A continuación se mostrarán los resultados del sistema de forma gráfica, donde se muestra el anafor con sus antecedentes y la solución ofrecida por el sistema.

Nótese que las frases son los resultados del análisis sintáctico, por lo que ya están procesadas, por ejemplo, las clíticas nominales, como *lo*, están separadas de los verbos para hacer más clara su comprensión, aunque según las reglas del español deban ir juntos.

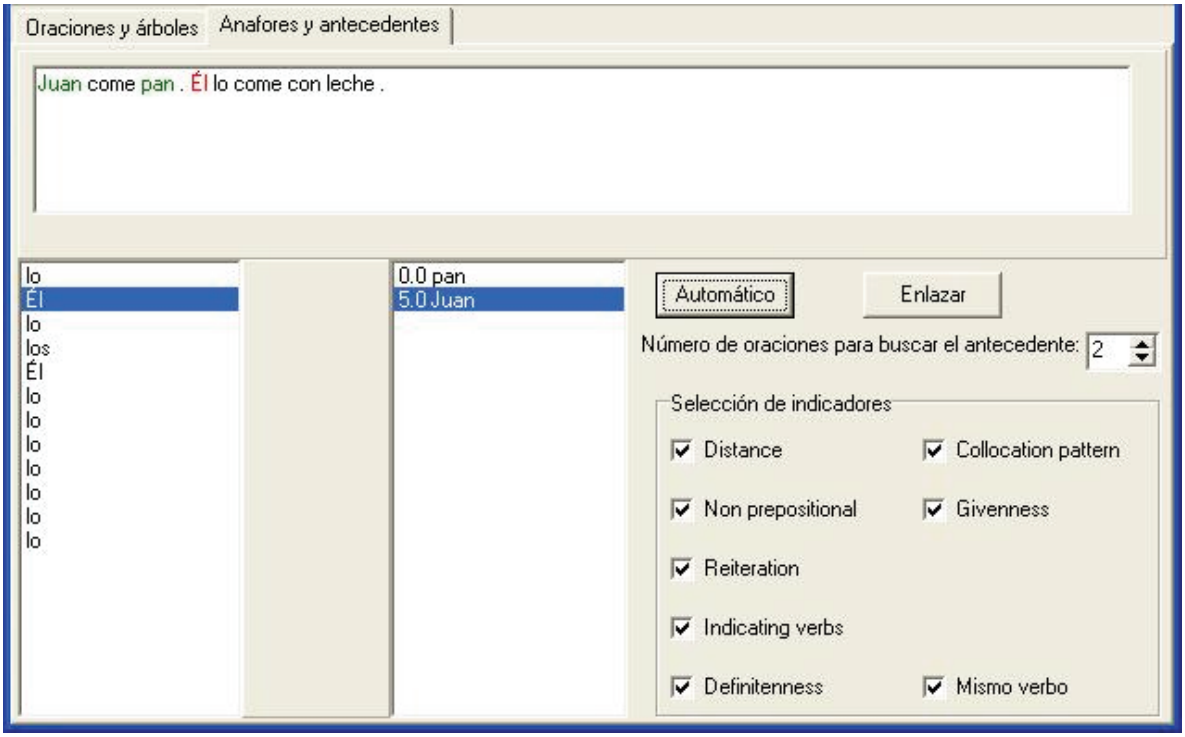


En esta imagen se muestran las oraciones

Juan come pan.

Él lo come con leche.

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

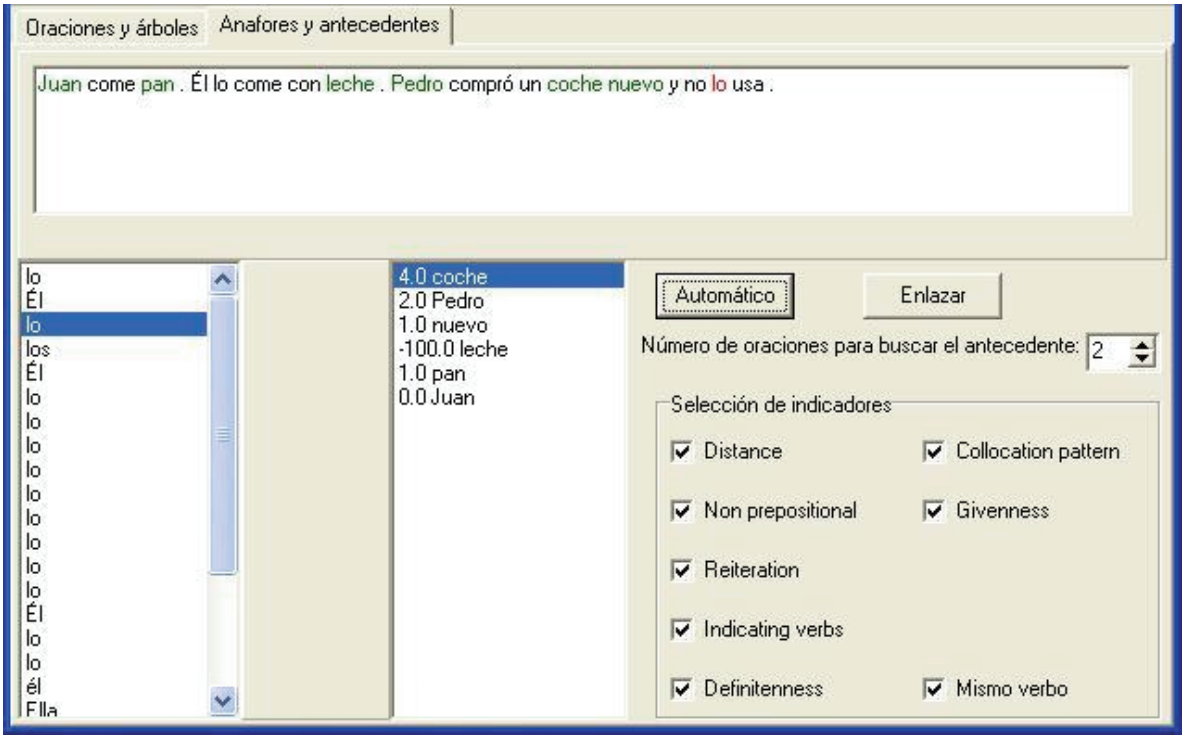


En esta imagen se muestran las oraciones

Juan come pan.

Él lo come con leche.

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



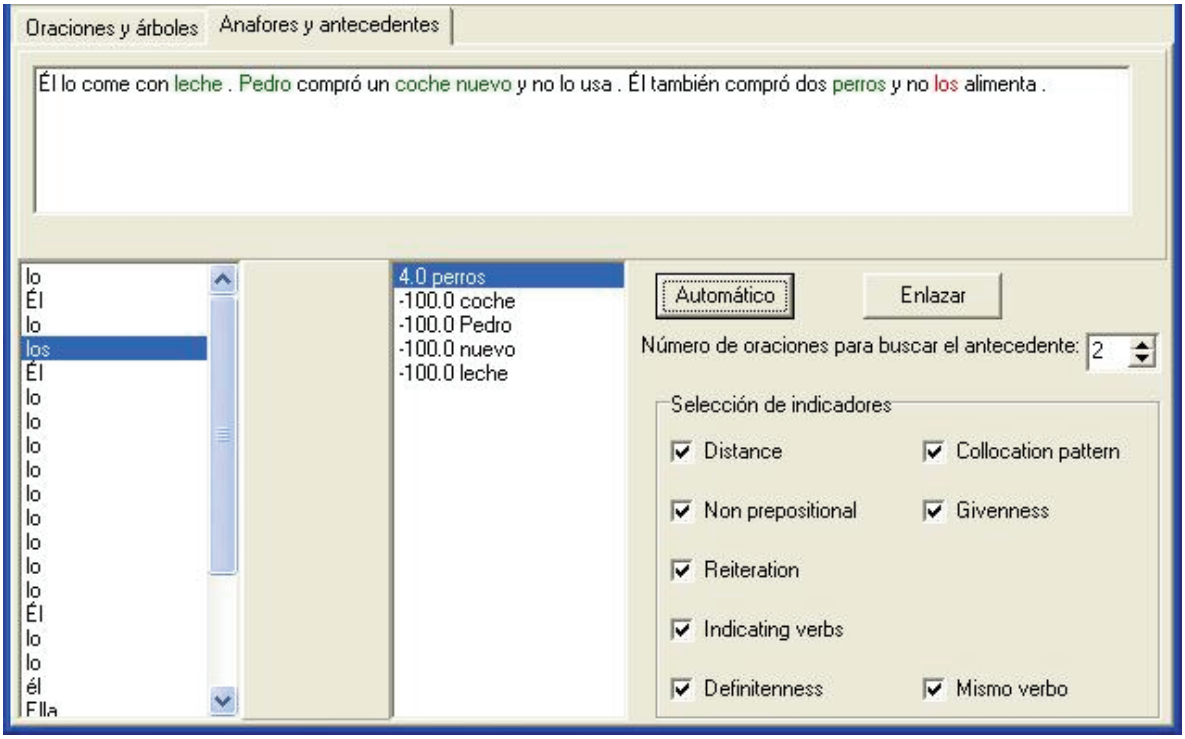
En esta imagen se muestran las oraciones

Juan come pan.

Él lo come con leche.

Pedro compró un coche nuevo y no lo usa.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



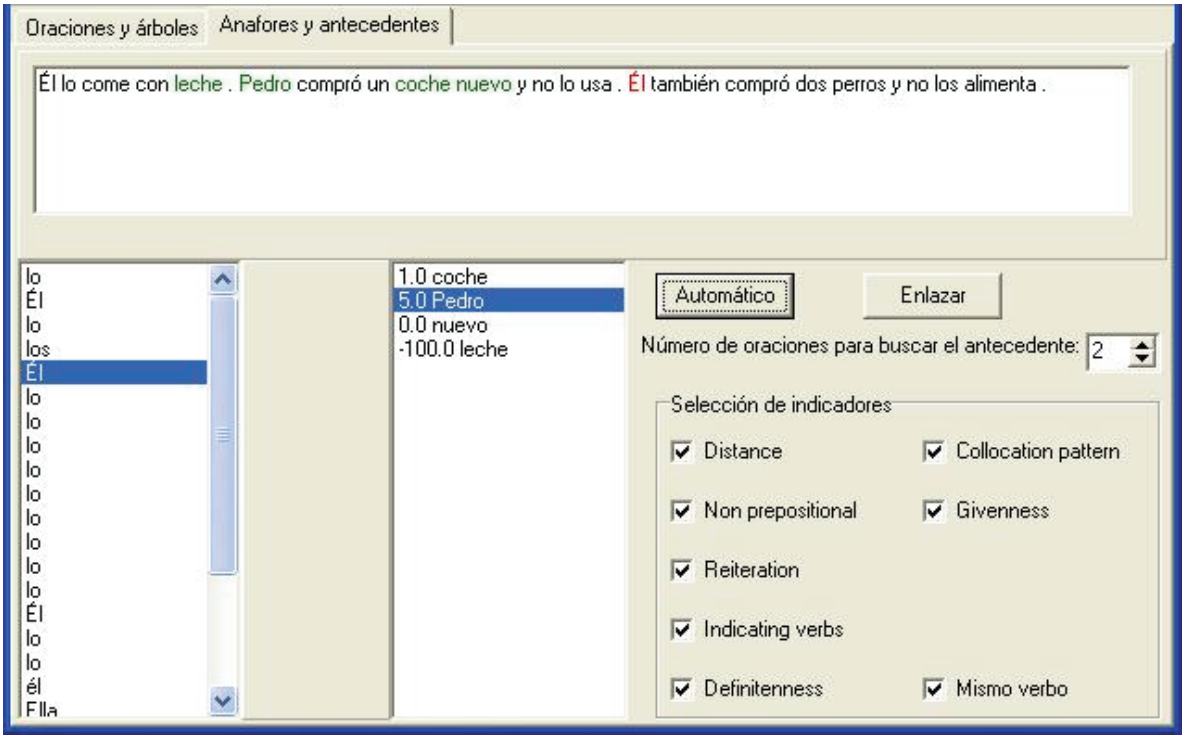
En esta imagen se muestran las oraciones

Él lo come con leche.

Pedro compró un coche nuevo y no lo usa.

Él también compró dos perros y no los alimenta.

El anafor a solucionar es el elemento *los* de la tercera oración, observándose que el sistema ofrece la solución correcta.



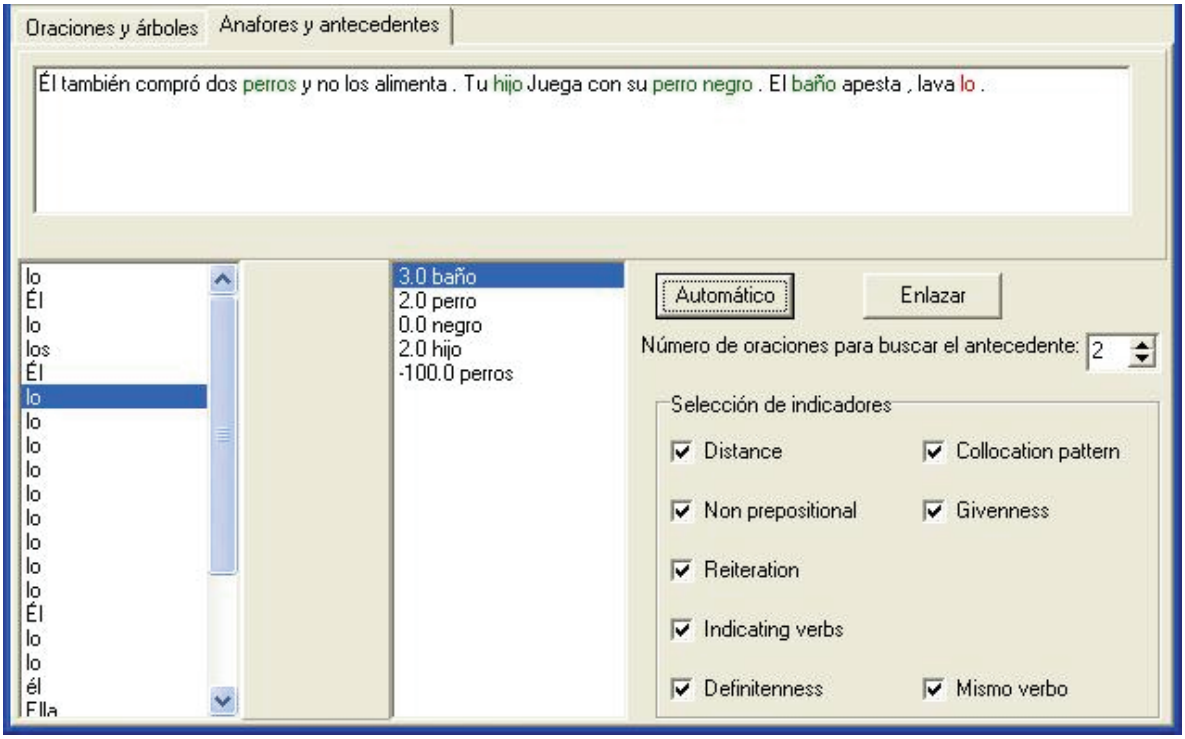
En esta imagen se muestran las oraciones

Él lo come con leche.

Pedro compró un coche nuevo y no lo usa.

Él también compró dos perros y no los alimenta.

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



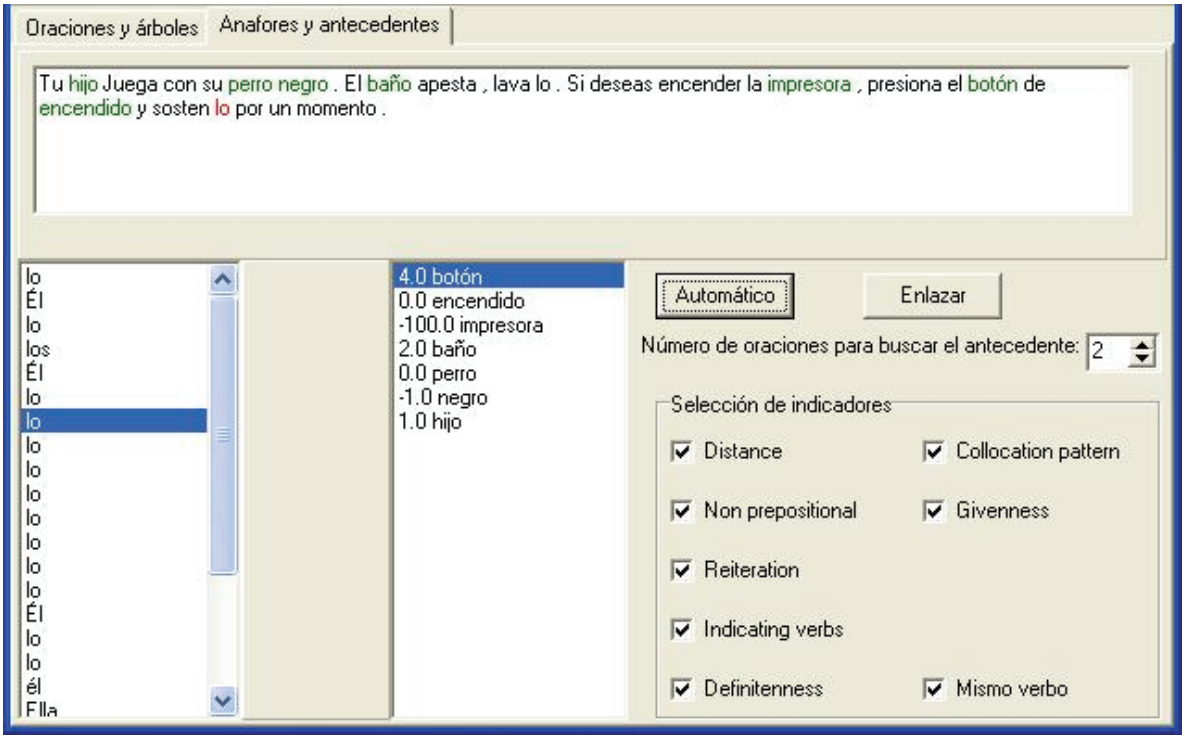
En esta imagen se muestran las oraciones

Él también compró dos perros y no los alimenta.

Tu hijo Juega con su perro negro.

El baño apesta, lava lo.

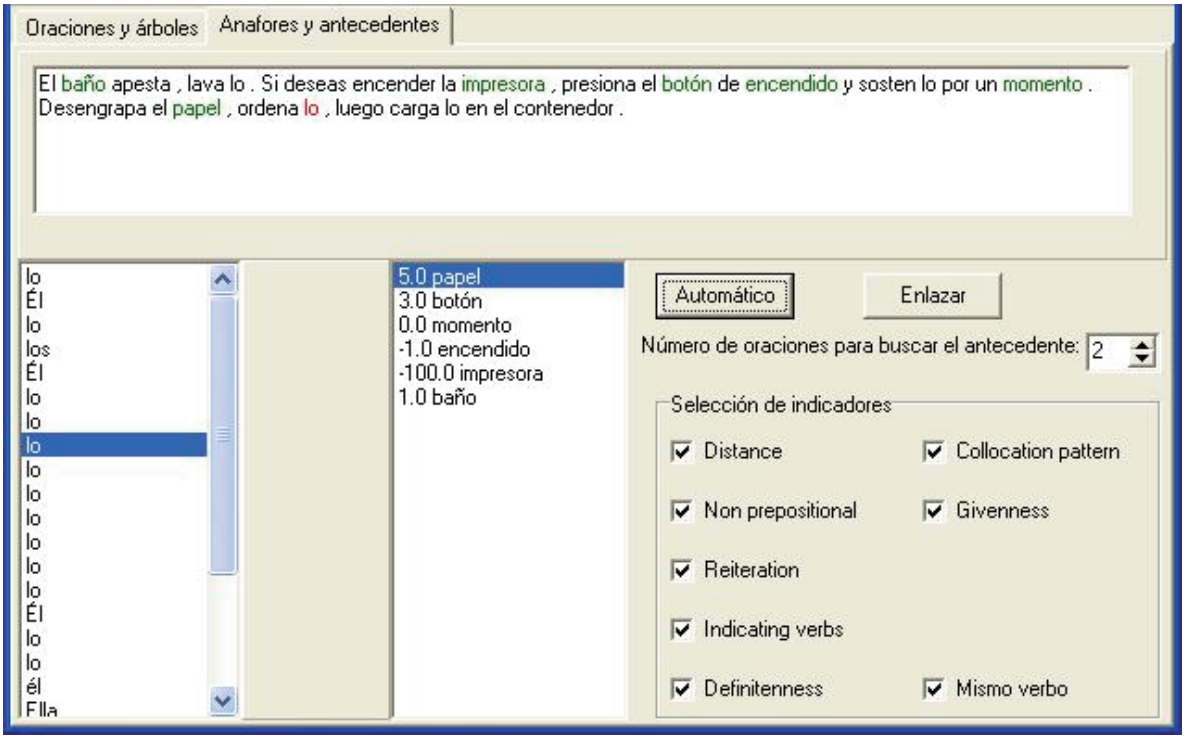
El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



En esta imagen se muestran las oraciones

> *Tu hijo Juega con su perro negro.*
>
> *El baño apesta, lava lo.*
>
> *Si deseas encender la impresora, presiona el botón de encendido y sosten lo por un momento.*

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



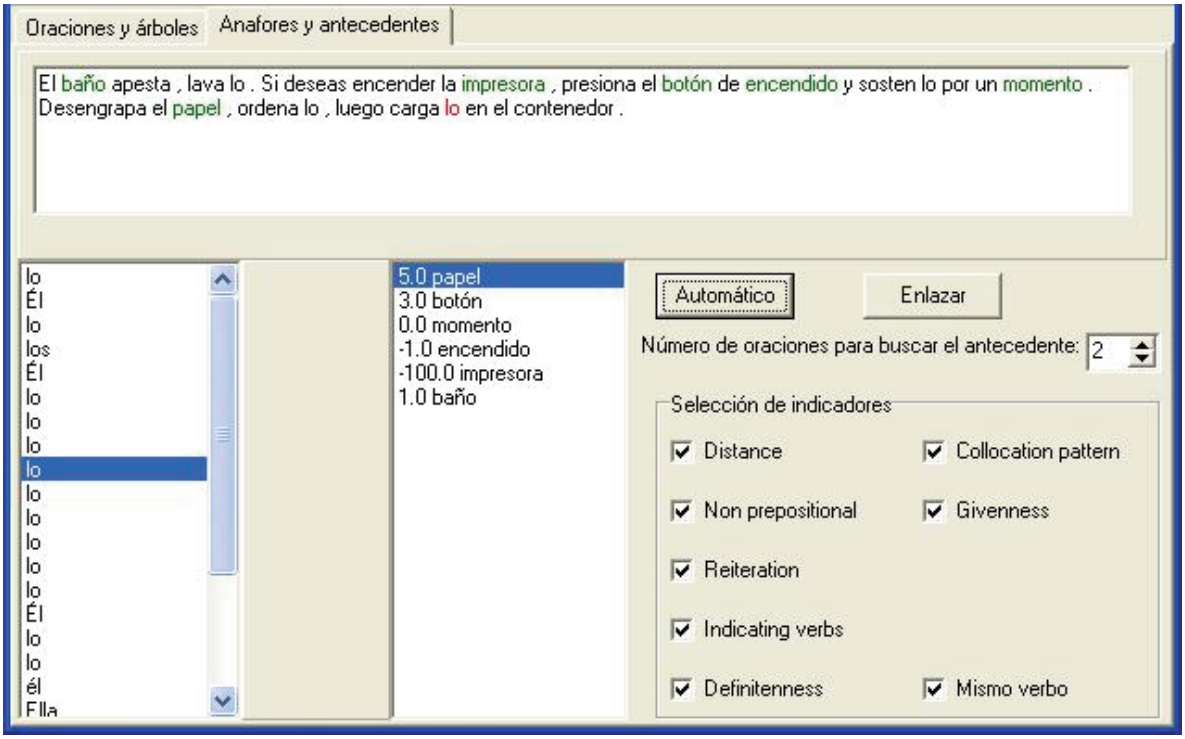
En esta imagen se muestran las oraciones

El baño apesta, lava lo.

Si deseas encender la impresora, presiona el botón de encendido y sosten lo por un momento.

Desengrapa el papel, ordena lo, luego carga lo en el contenedor.

El anafor a solucionar es el primer elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



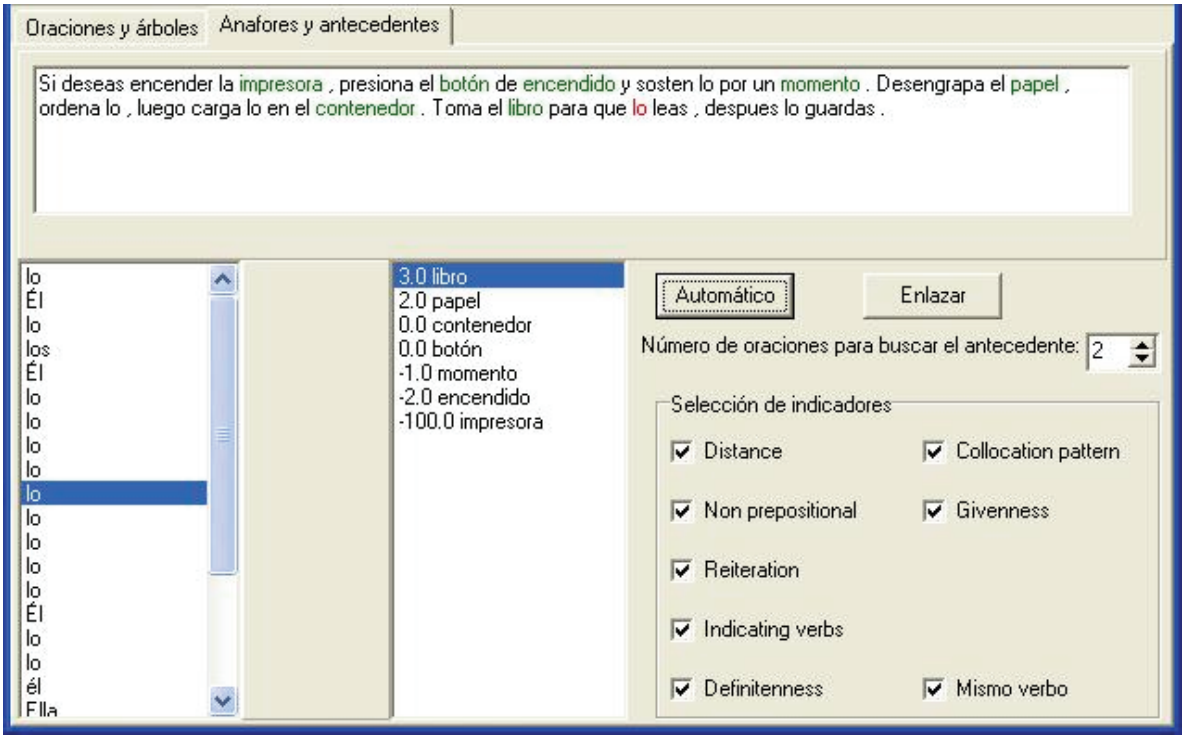
En esta imagen se muestran las oraciones

El baño apesta, lava lo.

Si deseas encender la impresora, presiona el botón de encendido y sosten lo por un momento.

Desengrapa el papel, ordena lo, luego carga lo en el contenedor.

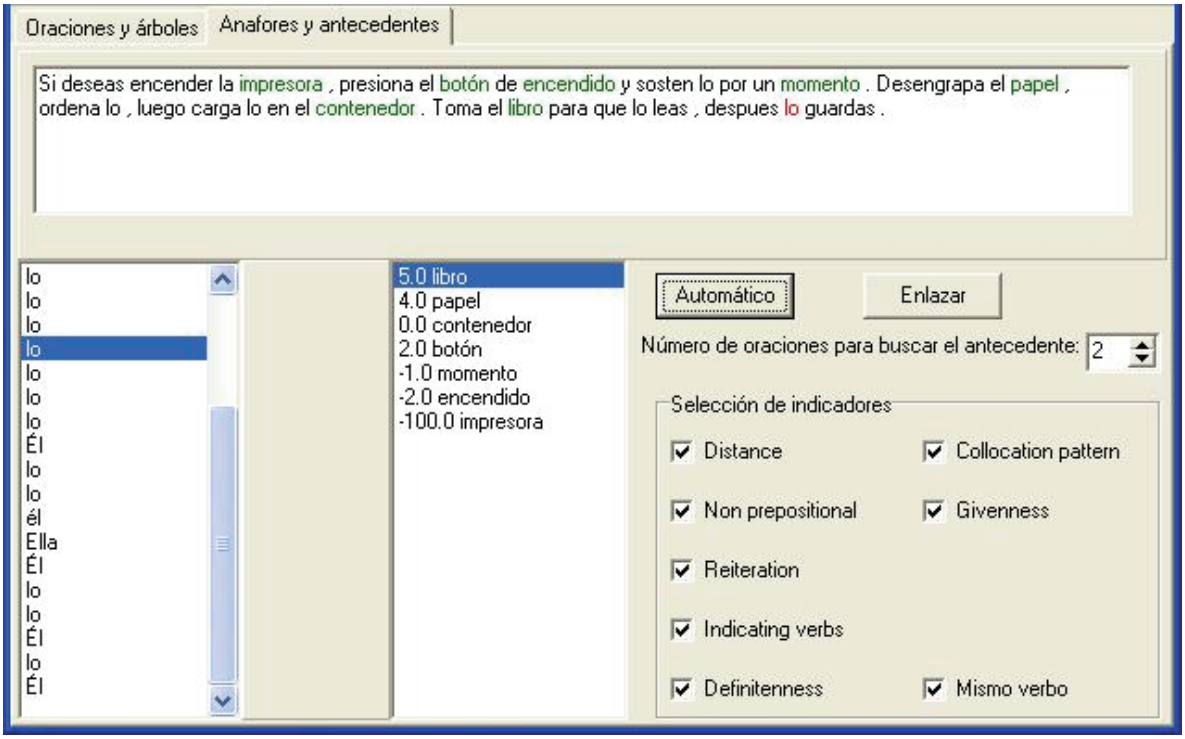
El anafor a solucionar es el segundo elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



En esta imagen se muestran las oraciones

> *Si deseas encender la impresora, presiona el botón de encendido y sosten lo por un momento.*
>
> *Desengrapa el papel, ordena lo, luego carga lo en el contenedor.*
>
> *Toma el libro para que lo leas, después lo guardas.*

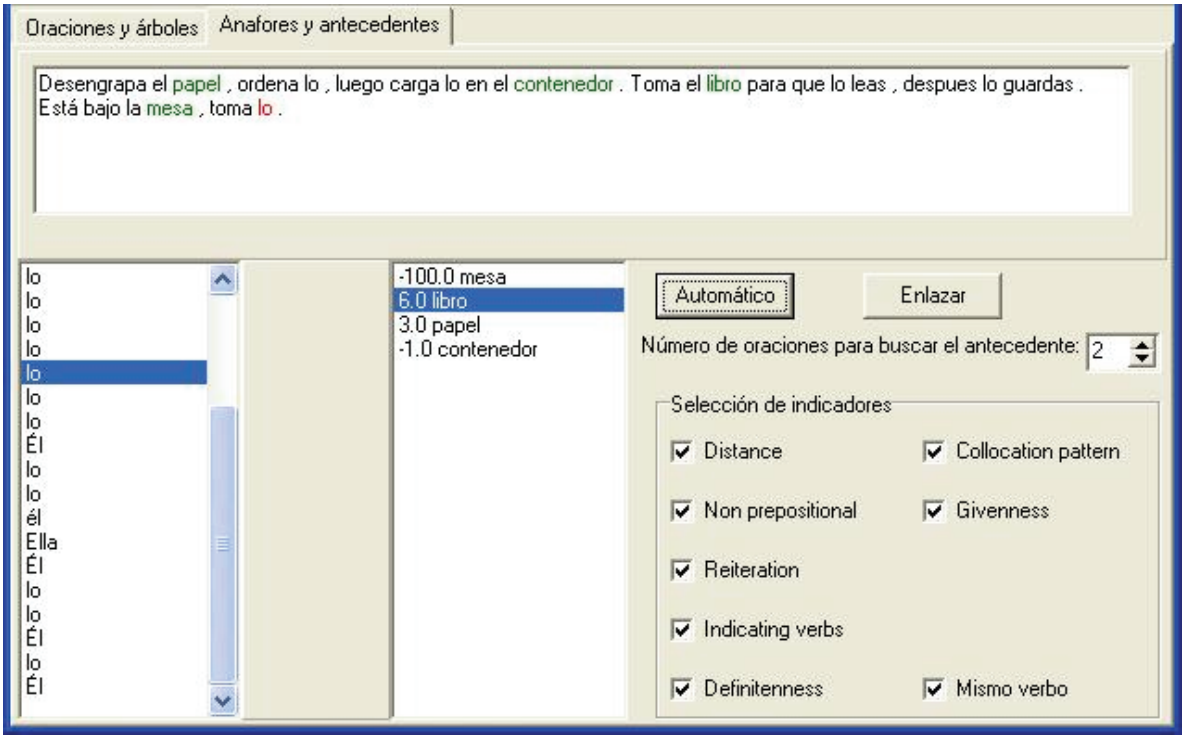
El anafor a solucionar es el primer elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



En esta imagen se muestran las oraciones

> *Si deseas encender la impresora, presiona el botón de encendido y sosten lo por un momento.*
>
> *Desengrapa el papel, ordena lo, luego carga lo en el contenedor.*
>
> *Toma el libro para que lo leas, después lo guardas.*

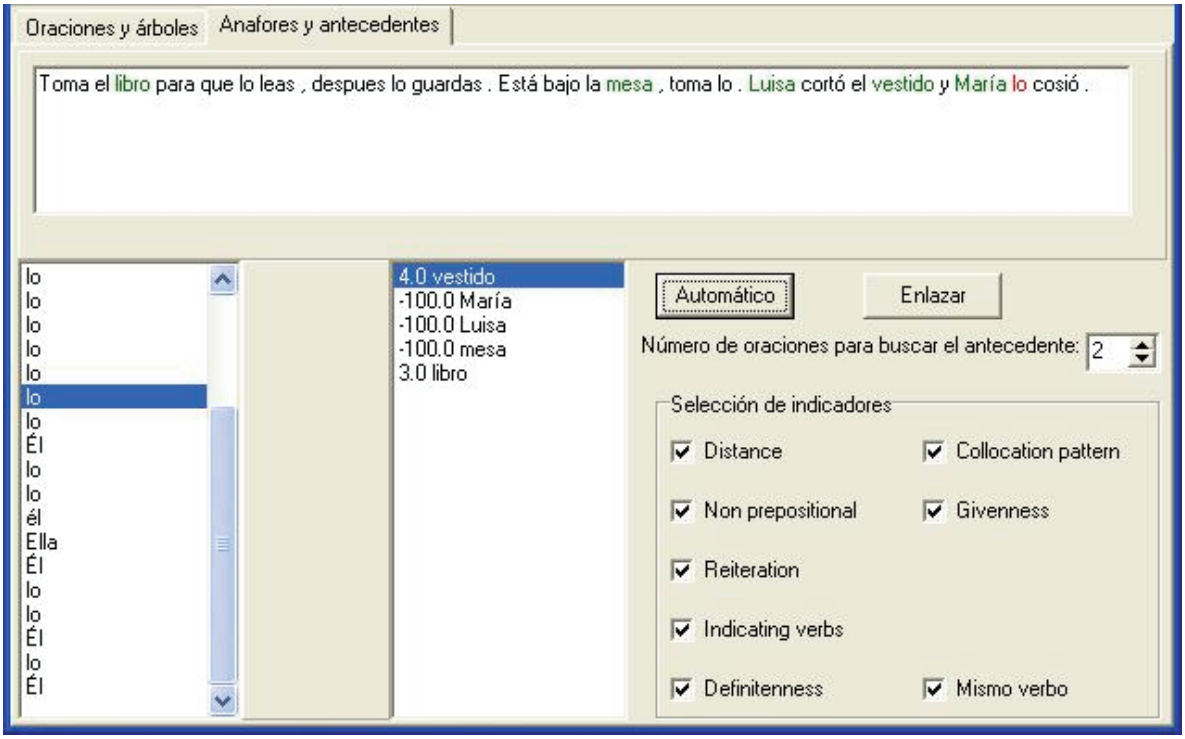
El anafor a solucionar es el segundo elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



En esta imagen se muestran las oraciones

> *Desengrapa el papel, ordena lo, luego carga lo en el contenedor.*
>
> *Toma el libro para que lo leas, después lo guardas.*
>
> *Está bajo la mesa, toma lo.*

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



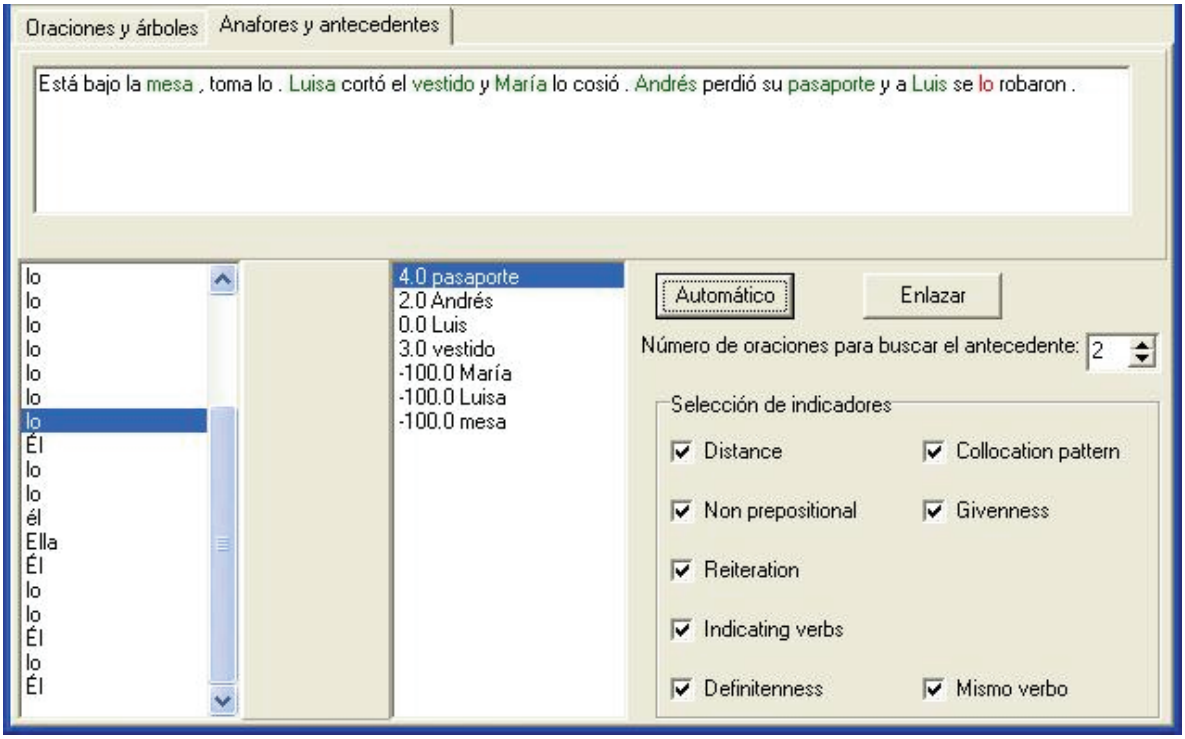
En esta imagen se muestran las oraciones

Toma el libro para que lo leas, despues lo guardas.

Está bajo la mesa, toma lo.

Luisa cortó el vestido y María lo cosió.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



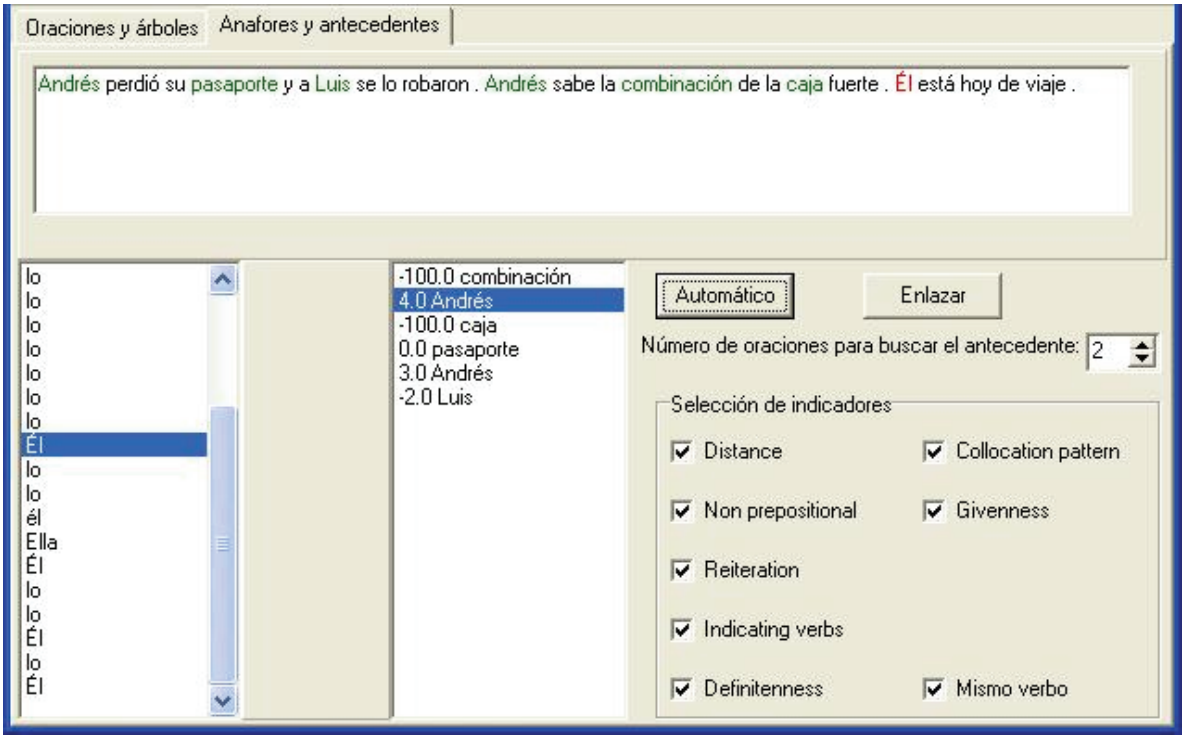
En esta imagen se muestran las oraciones

Está bajo la mesa, toma lo.

Luisa cortó el vestido y María lo cosió.

Andrés perdió su pasaporte y a Luis se lo robaron.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



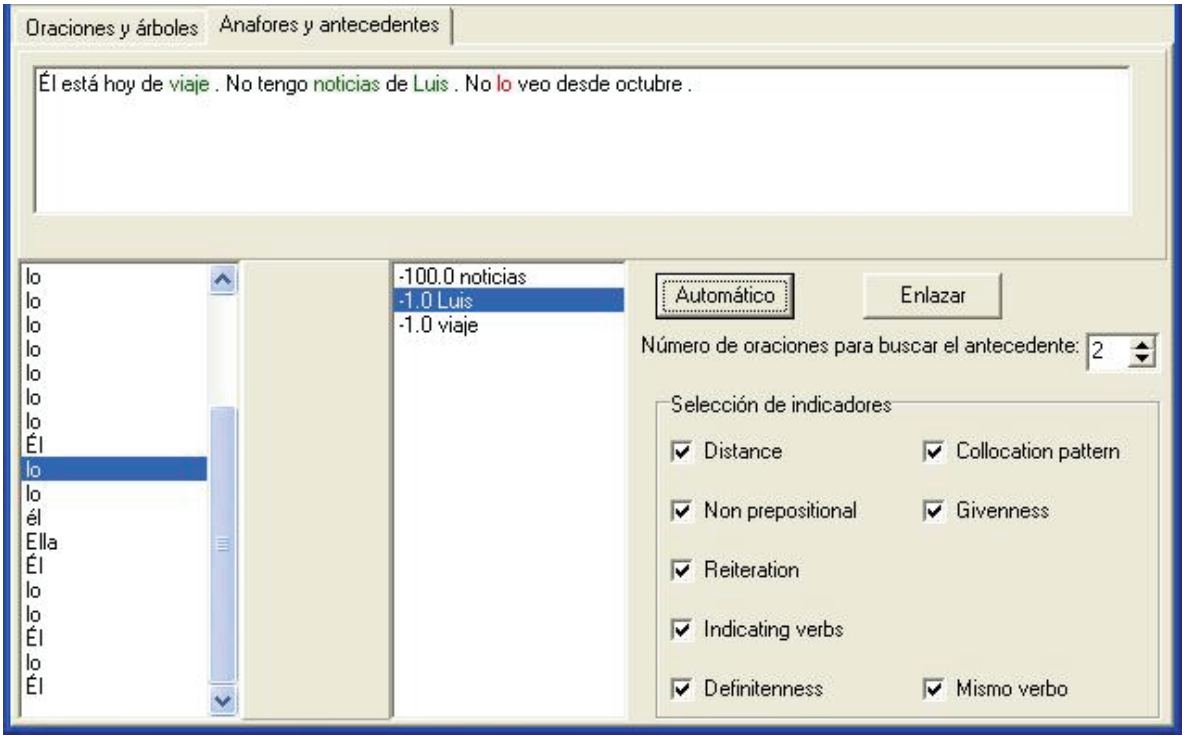
En esta imagen se muestran las oraciones

Andrés perdió su pasaporte y a Luis se lo robaron.

Andrés sabe la combinación de la caja fuerte.

Él está hoy de viaje.

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



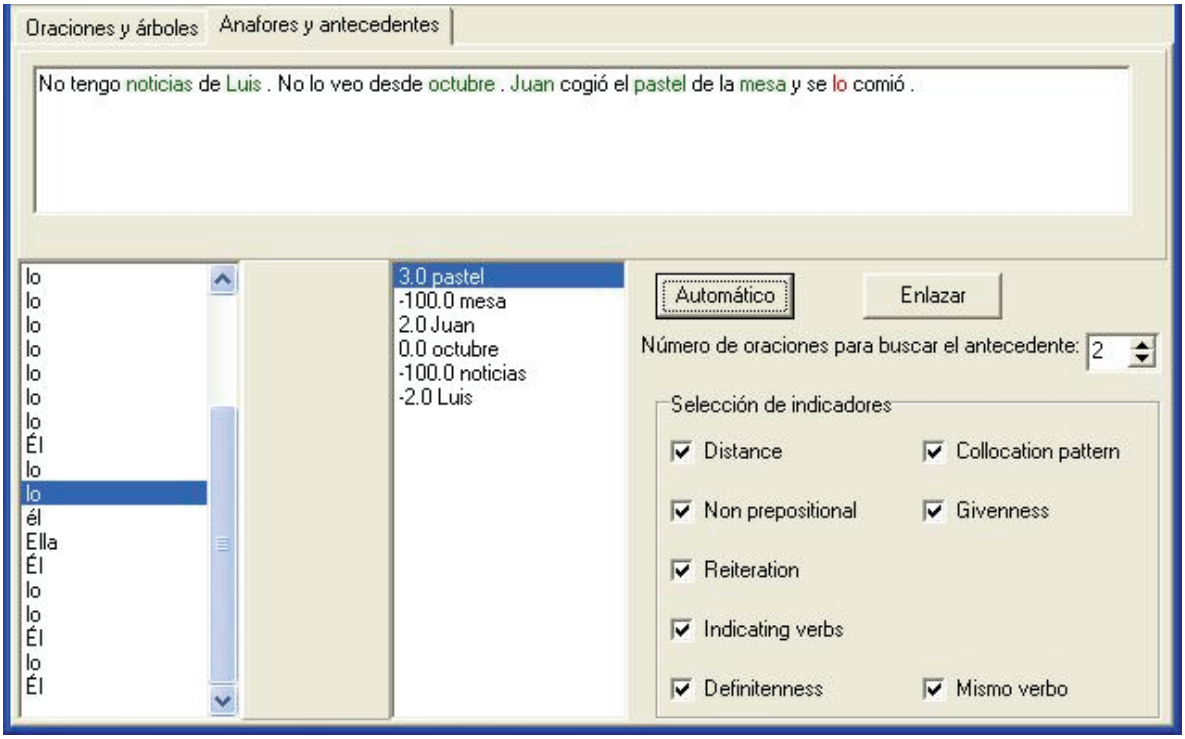
En esta imagen se muestran las oraciones

Él está hoy de viaje.

No tengo noticias de Luis.

No lo veo desde octubre.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



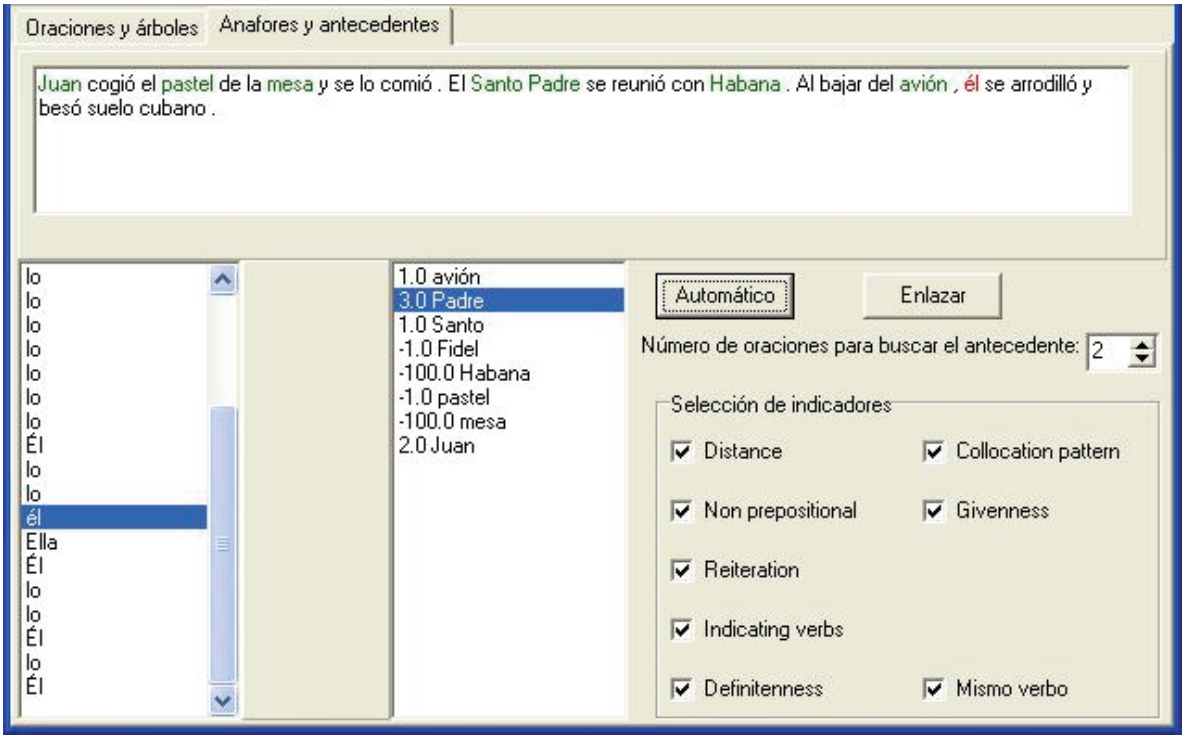
En esta imagen se muestran las oraciones

No tengo noticias de Luis.

No lo veo desde octubre.

Juan cogió el pastel de la mesa y se lo comió.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



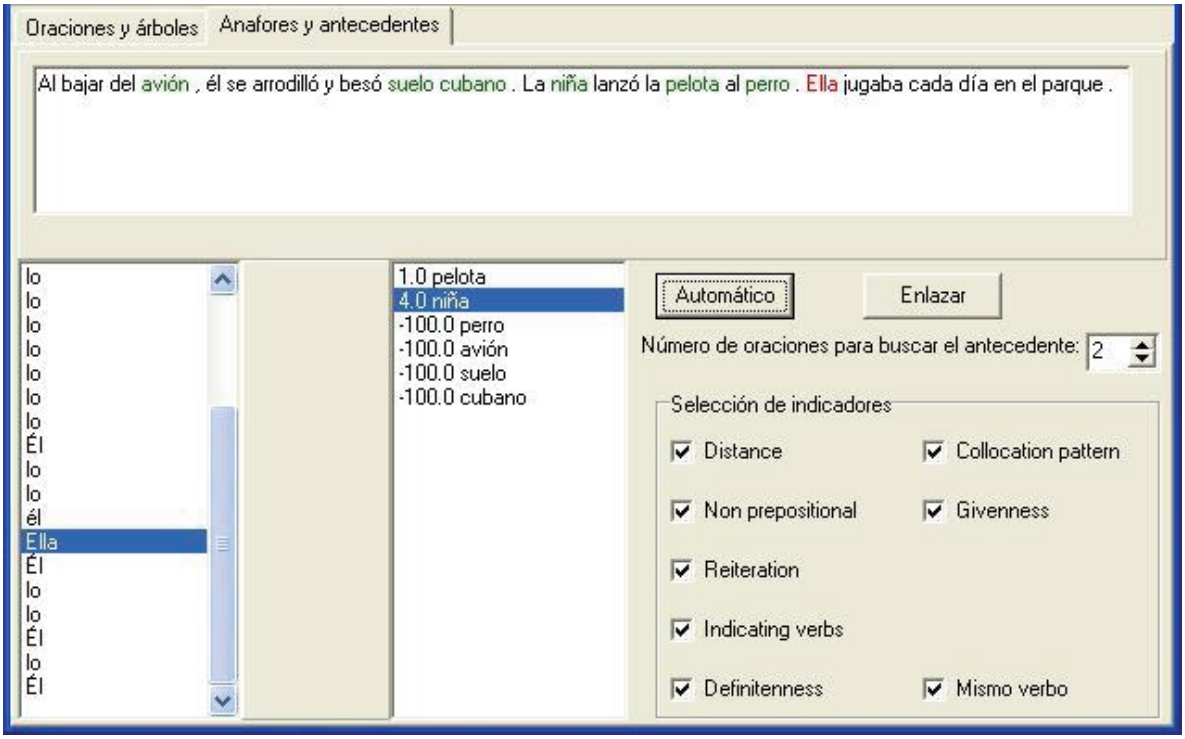
En esta imagen se muestran las oraciones

Juan cogió el pastel de la mesa y se lo comió.

El Santo Padre se reunió con Fidel en La Habana.

Al bajar del avión, él se arrodilló y besó suelo cubano.

El anafor a solucionar es el elemento *él* de la tercera oración, observándose que el sistema ofrece la solución correcta.



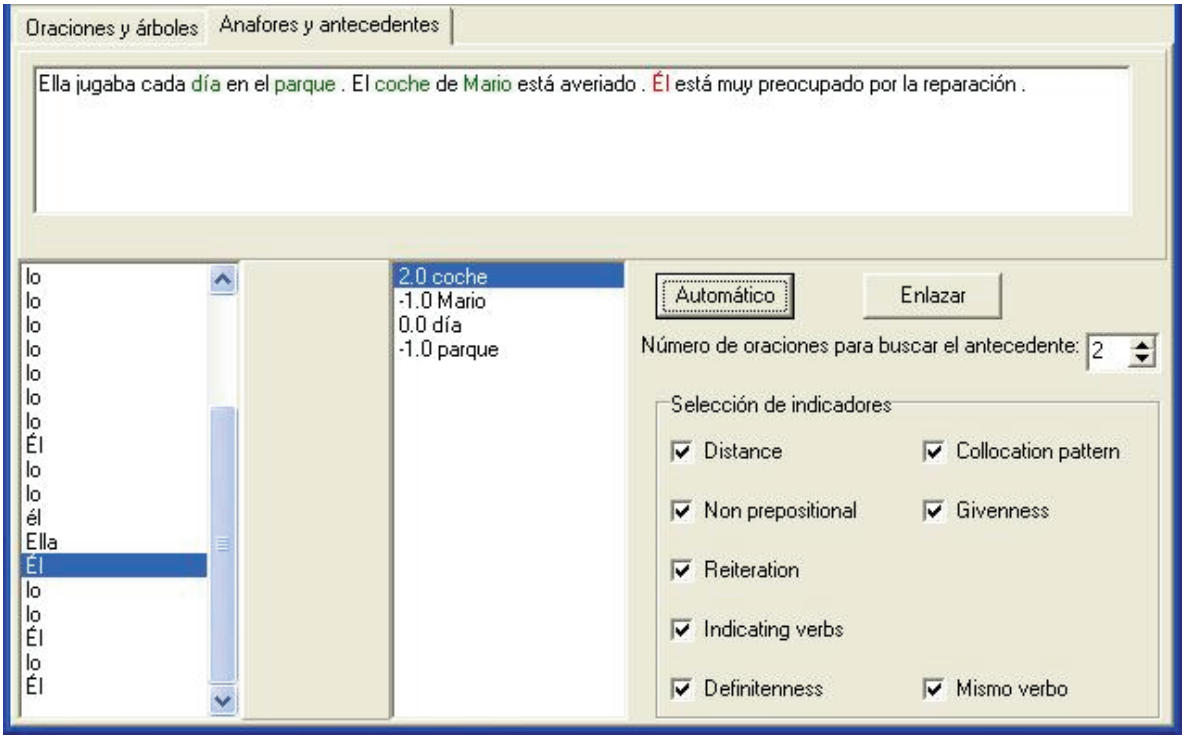
En esta imagen se muestran las oraciones

Al bajar del avión, él se arrodilló y besó suelo cubano.

La niña lanzó la pelota al perro.

Ella jugaba cada día en el parque.

El anafor a solucionar es el elemento *ella* de la tercera oración, observándose que el sistema ofrece la solución correcta.



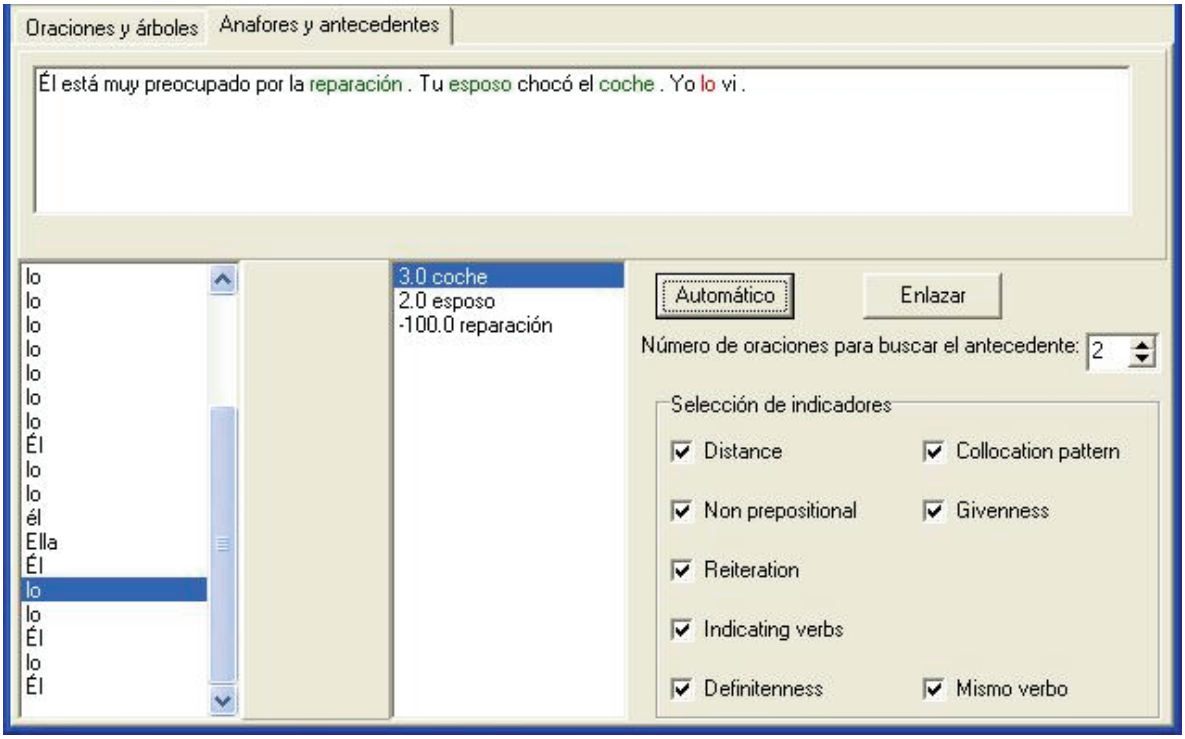
En esta imagen se muestran las oraciones

Ella jugaba cada día en el parque.

El coche de Mario está averiado.

Él está muy preocupado por la reparación.

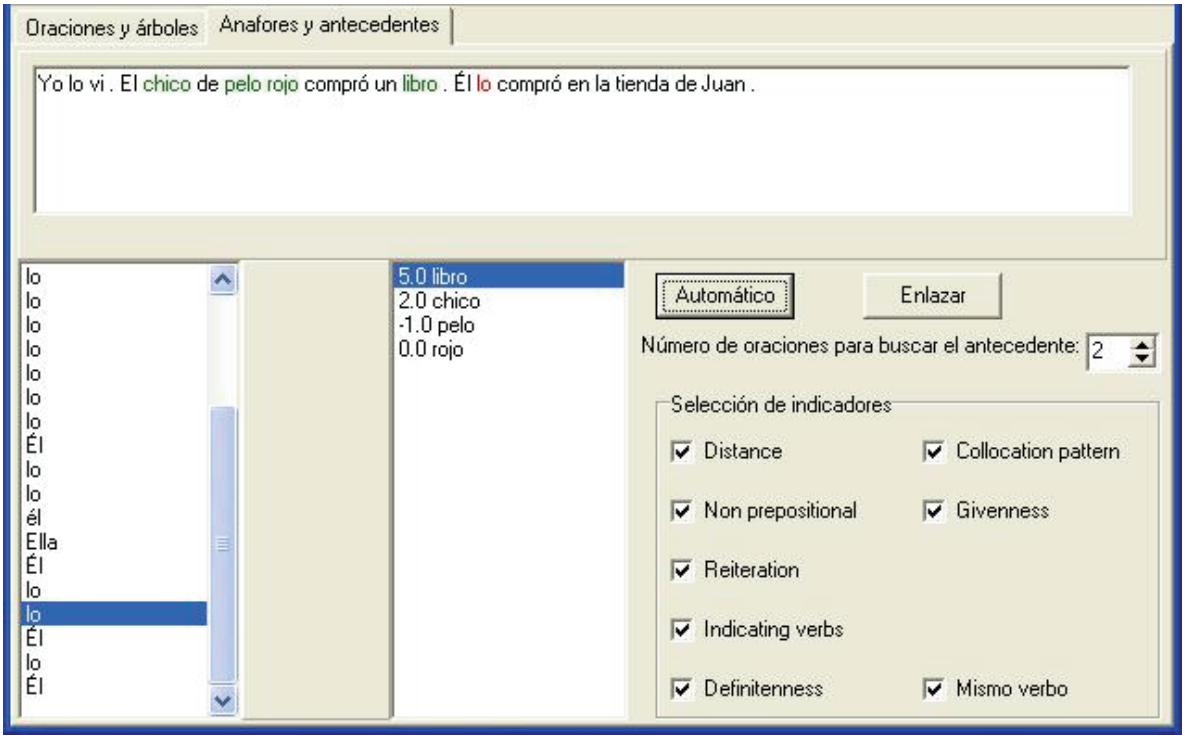
El anafor a solucionar es el elemento *Él* de la tercera oración, observándose que el sistema ofrece una solución incorrecta.



En esta imagen se muestran las oraciones

> *Él está muy preocupado por la reparación.*
>
> *Tu esposo chocó el coche.*
>
> *Yo lo vi.*

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece una solución incorrecta.



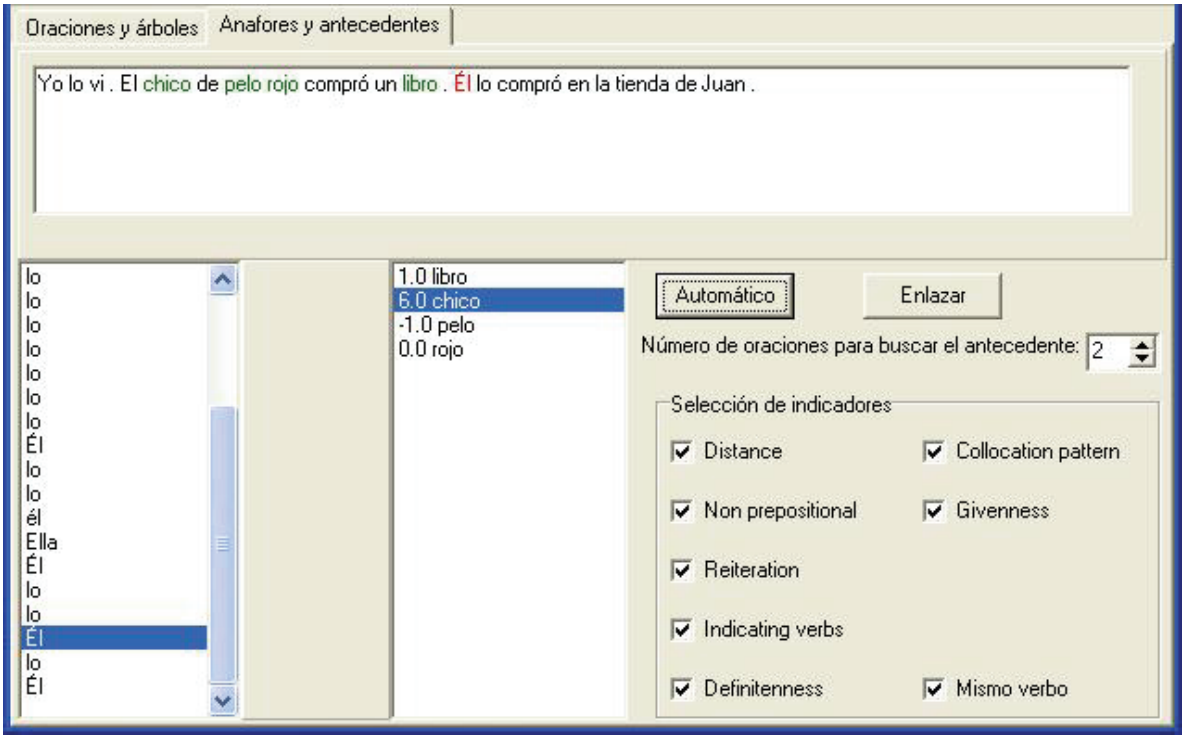
En esta imagen se muestran las oraciones

Yo lo vi.

El chico de pelo rojo compró un libro.

Él lo compró en la tienda de Juan.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



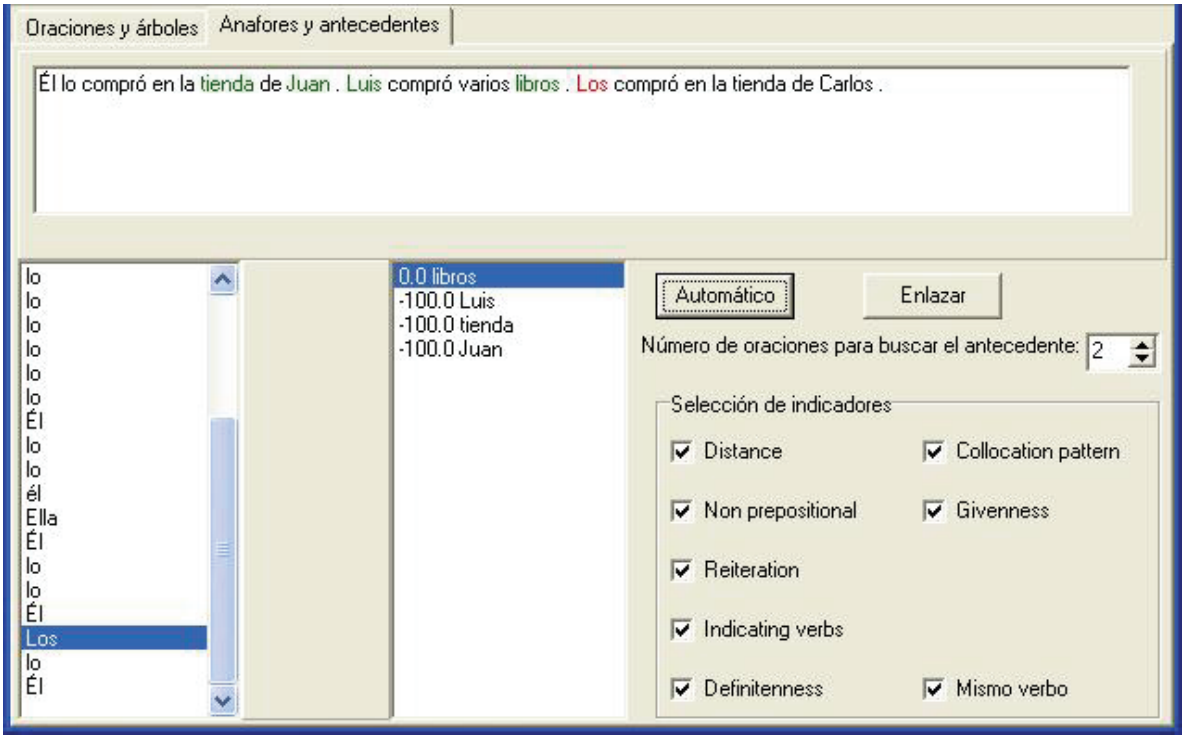
En esta imagen se muestran las oraciones

Yo lo vi.

El chico de pelo rojo compró un libro.

Él lo compró en la tienda de Juan.

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



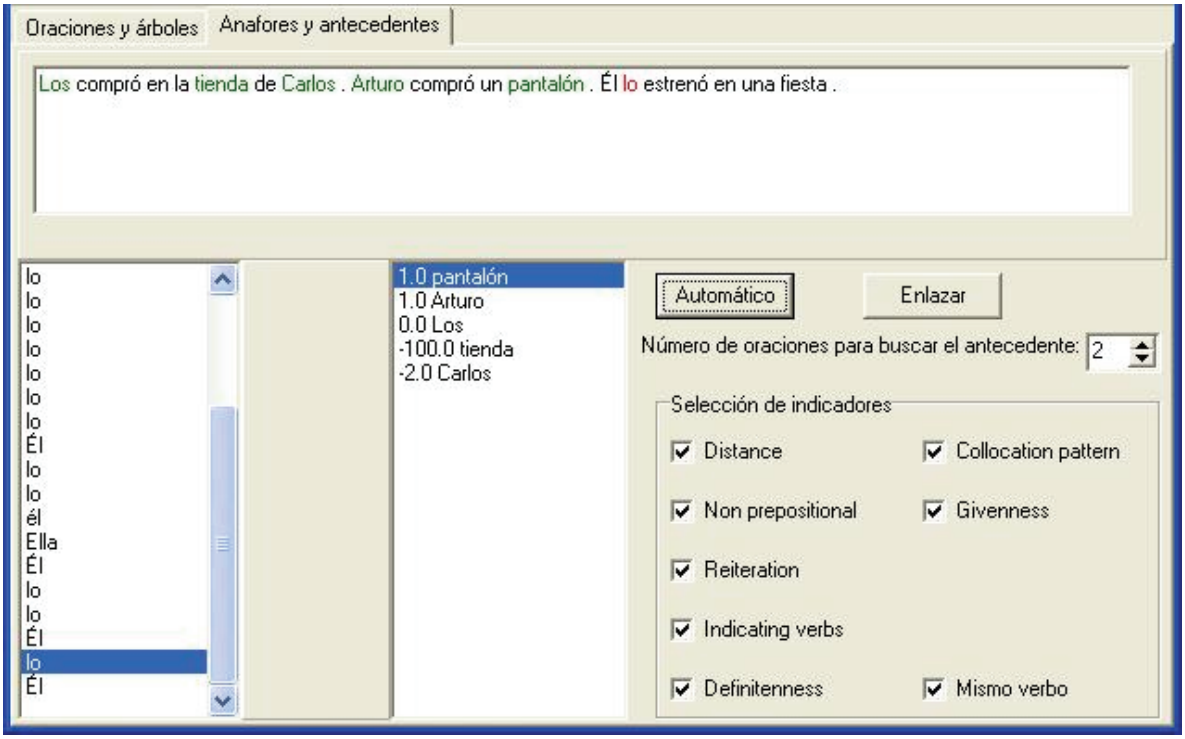
En esta imagen se muestran las oraciones

Él lo compró en la tienda de Juan.

Luis compró varios libros.

Los compró en la tienda de Carlos.

El anafor a solucionar es el elemento *Los* de la tercera oración, observándose que el sistema ofrece la solución correcta.



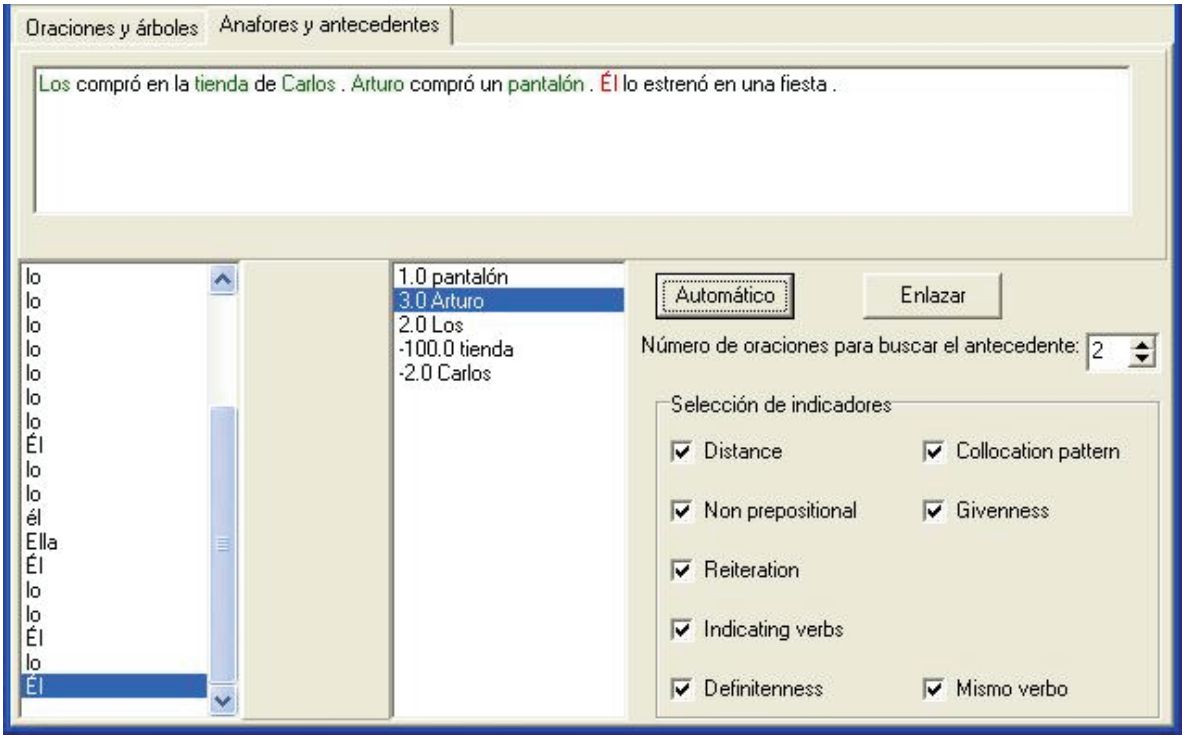
En esta imagen se muestran las oraciones

Los compró en la tienda de Carlos.

Arturo compró un pantalón.

Él lo estrenó en una fiesta.

El anafor a solucionar es el elemento *lo* de la tercera oración, observándose que el sistema ofrece la solución correcta.



En esta imagen se muestran las oraciones

Los compró en la tienda de Carlos.

Arturo compró un pantalón.

Él lo estrenó en una fiesta.

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

De lo anterior podemos observar que de 26 anáforas detectadas, el sistema resolvió correctamente 24, obteniendo un 92.30% de efectividad.

CAPÍTULO 6 CONCLUSIONES Y TRABAJO FUTURO

En este capítulo se dan las aportaciones, conclusiones y algunas sugerencias para trabajos futuros.

6.1 Aportaciones

- Se adaptó el método de R. Mitkov de resolución de anáfora para el español. Igualmente se hicieron las modificaciones relacionadas con el uso de un analizador sintáctico automático.
- Se implementó este método de resolución de la anáfora pronominal con las modificaciones pertinentes para el español.
- Se desarrolló el sistema que implementa el método en lenguaje C++.
- Se realizaron pruebas de funcionamiento del método usando textos en el español.
- El método tiene alta efectividad (92.30%) en los textos de prueba.

6.2 Conclusiones

Para la construcción de redes semánticas y generalmente para cualquier procesamiento automático a nivel profundo, es necesario resolver la anáfora en textos.

Se desarrolló la herramienta (el sistema) que permite hacer esta resolución de la anáfora pronominal para el español.

Para eso se implementó el método de R. Mitkov para el español con las modificaciones pertinentes. También una de las diferencias importantes es que en nuestro caso usamos el analizador sintáctico a diferencia del método original.

Se hicieron pruebas que dieron 92.30% de precisión sobre un texto pequeño que consiste en 31 oraciones.

6.3 Trabajos futuros

Para trabajos futuros se consideraron los siguientes puntos:

- Hacer pruebas sobre una colección de textos más grande.
- Dar la posibilidad de evaluar las pruebas de manera automática.
- Comparar los resultados obtenidos con algún otro método de resolución de anáforas.

BIBLIOGRAFÍA

1. Allen, J. Natural Language Understanding. The Benjamin / Cummings Publishing Company, Inc., 1995.

2. Alshawi, H. Memory and Context for Language Interpretation. Cambridge University Press, 1987.

3. Amores, J.G. A Lexical-Functional Grammar Machine Translation System for Medical Abstracts. Tesis doctoral. Universidad de Sevilla, 1992.

4. Ariel, M. Referring and accessibility. Journal of Linguistics,24 (1990). pp 65-87.

5. Azzam, S. An Algorithm to Co-Ordinate Anaphora Resolution and PPS Disambiguation Process. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL'95 (1995).

6. Azzam, S. Anaphors, pps and disambiguation process for conceptual analysis. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95 (San Mateo, 1995).

7. Azzam, S. Computation of Ambiguities (Anaphors and PPs) in NL texts. CLAM: The prototype. Ph.D. thesis. Paris Sorbonne University, 1995.

8. Baldwin, B. CogNIAC: a discourse processing engine. Ph.D. dissertation. University of Pennsylvania, Department of Computer and Information Science, 1995.

9. Baldwin, Breck (1997). "CogNIAC: high precision coreference with limited knowledge and linguistic resources", en Proceedings of ACL/EACL

workshop on Operational factors in practical, robust anaphora resolution. 38-45.

10. Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaparkhi, A., Rosenzweig, J., Sarkar, A., and Srinivas. University of Pennsylvania: Description of the University of Pennsylvania system used for MUC-6. In Proceedings of the Sixth Message Understanding Conference (1995).

11. Bobrow. The high-school algebra problem answering system STUDENT with natural language input, 1964.(*)

12. Bobrow, D., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. Gus, a frame driven dialog system. Artificial Intelligence, 8 (1977). pp. 155-173.

13. Botley, S. Comparing demonstrative features in three written English genres. In Proceedings of the Discourse Anaphora And Resolution Colloquium, DAARC96 (1996).

14. Brennan, S., Friedman, M., and Pollard, C. A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, ACL'87 (1987). pp. 155-162.

15. Brown, G., and Yule, G. Discourse Analysis. Cambridge University Press, 1983.

16. Carbonell, James G. & Ralf D. Brown. 1988. "Anaphora resolution: a multi-strategy approach". Proceedings of the 12. International Conference on Computational Linguistics (COLING'88), Vol.I, 96-101, Budapest, Hungria.

17. Carter, D. Control issues in anaphor resolution. Journal of Semantics,7 (1990). pp. 435-454.

8. Carter, D. Interpreting Anaphors in Natural Language. Ellis Horwood, Ed., Chichester, UK, 1987.

9. Charniak, E. Toward a Model of Children's Story Comprehension. Cambridge, Mass: MIT A.I. Lab TR-266. 1972.

20. Chierchia, G. Anaphora and Dynamic Binding. Linguistics and Philosophy,15 (1992). pp. 111-183.

21. Chomsky, N. Knowledge of Language. Praeger, New York. 1986.

22. Chomsky, N. The New Sintaxis: Government and Binding Theory, 1988.

23. Chomsky, N. Lectures on Government and Binding. Foris Publications, Dordrecht, 1981.

24. Collins, M. A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL'96 (1996).

25. Connoly, D., Burger, J., and Day, D. A machine learning approach to anaphoric reference. In Proceedings of the International Conference on New Methods in Language Processing (UMIST, Manchester, 1994).

26. Covington, M. Natural Language Processing for Prolog Programmers. Prentice Hall, 1994.

27. Cullingford, R. E. Inside Computer Understanding: Five Programs Plus Miniatures. Lawrence Erlbaum, Hillsdale, NJ. In R. C. Schank and C. K. Riesbeck, Eds. 1981.

28. Dagan, Ido y Alon Itai (1991). "A statistical filter for resolving pronoun references", Artificial Intelligence and Computer Vision, 125-135.

29. Declerck, T. Dealing with Cross-Sentential Anaphora Resolution in ALEP. In Proceedings of the 16th International Conference on Computational Linguistics, COLING'96 (Copenhagen, Denmark, 1996). Vol. I, pp. 280-285.

30. Di Eugenio, B. The discourse functions of Italian subjects: a centering approach. In Proceedings of the 16th International Conference on Computational Linguistics, COLING'96 (Copenhagen, Denmark, 1996). Vol. I, pp. 352-357.

31. Ersan, E., and Akman, V. *Focusing for pronoun resolution in english discourse: an implementation.* BILKENT UNIVERSITY. Department of Computer Engineering and Information Science. Technical Report BU-CEIS-94-29, 1994.

32. Fernández, Antonio (1998). "Aproximación computacional al tratamiento de la anáfora pronominal y de tipo adjetivo mediante gramáticas de unificación de huecos", Tesis Doctoral. Universidad de Alicante, España.

33. Fligelstone, S. Developing a scheme for annotating text to show anaphoric relations. New directions in English language corpora: methodology, results, software development,9 (1992).

34. Franchini, E. Las condiciones gramaticales de la coordinación copulativa en español. Romanica Helvetica. Vol. 102. Francke Verlag Bern, 1986.

35. Gaizauskas, R., and Humpreys, K. Quantitative evaluation of coreference algorithms in an information extraction system. In Corpus-based and computational approaches to discourse anaphora. S.P. Botley and A.M. McEnery, Eds., UCL Press. 1997.

36. Grober, Beardsley and Caramazza. The implicit verb causality or causal valence, 1978.

37. Grosz, B. Focusing and description in Natural Language Dialogues. In Elements of Discourse Understanding. Cambridge: Cambridge University Press. 1981.

38. Grosz, B. The representation and use of focus in a system for understanding dialogs. In In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, IJCAI (Cambridge, MA., 1977).

39. Grosz, B., and Sidner, C. Attentions, intentions and the structure of discourse. Computational Linguistics,12. 3 (1986). pp. 175-204.

40. Grosz, B., Joshi, A., and Weinstein, S. Providing a unified account of definite noun phrases in discourse. In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, ACL'83 (1983). pp. 44-50.

41. Grosz, B., Aravind, J., and Weinstein, S. Centering: A Framework for Modeling the Local Coherence of Discourse. Computational Linguistics,21, 2 (1995). pp. 203-225.

42. Handbook on computational linguistics. Oxford Press, 2005.

43. Haegeman, Liliane. 1994. Introduction to Government and Binding Theory. Oxford: Blackwell.

44. Hardt, D. Centering in Dynamic Semantics. In Proceedings of the16th International Conference on Computational Linguistics, COLING'96 (Copenhagen, Denmark, 1996). Vol. I, pp. 519-524.

45. Heim, Y. The semantics of definite and indefinite noun phrases. Dissertation, University of Massachusetts. Amherst, 1982.

46. Hirst, Graeme. Anaphora in Natural Language Understanding. Berlin: Springer-Verlag, 1981.

47. Hobbs, J. Pronoun Resolution. Nueva York: Research Report # 76-1, Department of Computer Sciences. City College, City University of New York, 1976.

48. Hobbs, Jerry R. (1978) "Resolving pronoun references". Lingua, 44, 311-338.

49. Ingria, Robert J.P. & David Stallard. 1989. "A computational mechanism for pronominal reference". Proceedings of the 27th Annual Meeting of the ACL, 262-271, Vancouver, British Columbia.

50. Kameyama, M. A Property-Sharing Constraint in Centering. In Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, ACL'86 (1986). pp. 200-206.

51. Kameyama, M. Intrasentential Centering: A case Study. In Centering in Discourse. E. Prince, A. Joshi, and L. Walker, Eds., Oxford University Press. 1997.

52. Kennedy, Christopher y Branimir Boguraev (1996). "Anaphora for everyone: pronominal anaphora resolution without a parser", en Proceedings of 16th International Conference on Computational Linguitics. vol. I, 113-118.

53. Lappin, Shalom y Herbert Leass (1994). "An algorithm for pronominal anaphora resolution". Computational Linguistics. 20(4), 535-561.

54. Lappin, S., and McCord, M. A syntactic filter on pronominal anaphora in Slot Grammar. In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, ACL'90 (1990). pp. 135-142.

55. Lappin, S., and McCord, M. Anaphora resolution in Slot Grammar. Computational Linguistics,16. 4 (December 1990).

56. Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., and Shepherd, M. Cyc: Toward programs with common sense. Communications of the ACM,33, 8 (1990). pp. 30- 49.

57. Luperfoy, S. Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions. Ph.D. thesis. Department of Linguistics, The University of Texas at Austin, Austin, TX, 1991.

58. McEnery, Tony, and Andrew Wilson. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

59. McEnery, A., Tanaka, I., and Botley, S. Corpus annotation and reference resolution. In Proceedings of ACL/ EACL workshop on Operational factors in practical, robust anaphor resolution (Madrid, 1997).

60. Meya, M., and Hubert, W. Lingüistica computacional. Barcelona, Teide. 1986.

61. Mitkov, R. An integrated model for anaphora resolution. In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94 (Kyoto, Japan, 1994).

62. Mitkov, R. A new approach for tracking center. In Proceedings of the International Conference New Methods in Language Processing (Manchester, 1994).

63. Mitkov, Ruslan. 1995. Anaphora resolution in Natural Language Processing and Machine Translation. Working paper. Saarbrücken: IAI

64. Mitkov, Ruslan. 1995. "An uncertainty reasoning approach for anaphora resolution". Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95), 149- 154, Seul, Korea.

65. Mitkov, R. Two engines are better than one: generating more power and confidence in the search for the antecedent. In Proceedings of the Recent Advances in Natural Language Resolution, RANLP (1995).

66. Mitkov, Ruslan. 1996. "Anaphora resolution: a combination of linguistic and statistical approaches". Proceedings of the Discourse Anaphora and Anaphor Resolution (DAARC'96). Lancaster, UK.

67. Mitkov, Ruslan. 1997. "Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches" Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution., 14-21. Madrid, España.

68. Mitkov, Ruslan. How far are we from (semi-) automatic annotation of anaphoric links in corpora. In Proceedings of ACL/ EACL workshop on Operational factors in practical, robust anaphor resolution (Madrid, 1997).

69. Mitkov, R. Pronoun resolution: the practical alternative. In Corpus-based and Computational Approaches to Discourse Anaphora. S. Botley and T. McEnery, Eds., Univ. College London Press. 1997.

70. Mitkov, R., and Stys, M. 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).

71. Mitkov, Ruslan. 1998. "Robust pronoun resolution with limited knowledge". Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference. Montreal, Canada.

72. Mollá, D. On searching the antecedent of an anaphora. Department of Linguistics. Edinburgh, 1992.

73. Muskens, R. Combining montague semantics and discourse representation. Linguistics and Philosophy, 1996.

74. Nasukawa, T. Robust method of pronoun resolution using full-text information. In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94 (1994).

75. Okumura Manabu, Tamura Kouji. Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory. In Proceedings of the 16th International Conference on Computational Linguistics, COLING'96 (Copenhagen, Denmark, 1996). Vol. II, pp. 871-876.

76. Partee, B. H. Opacity, coreference, and pronouns. In Semantics of Natural Language. D. Davidson and G. Harman, Eds., D. Reidel Publishing Company, Dordrecht, Holland, 1972.

77. Palomar, Manuel, Antonio Fernández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Zaiz-Noeda y Rafael Muñoz (2001), "An Algorithm for Anaphora Resolution in Spanish Texts", Computacional Linguistics, 27(4), 545-567.

78. Peral Cortés, Jesús. Resolución y generación de la anáfora pronominal en español e inglés en un sistema interlingua de Traducción Automática. Tesis doctoral. Universidad de Alicante, 2001

79. Poesio, M., Vieira, R., and Teufel, S. Resolving bridging references in unrestricted text. In Proceedings of ACL/ EACL workshop on Operational factors in practical, robust anaphor resolution (Madrid, 1997).

80. Preuß S., B. Schmitz, C. Hauenschild & C. Umbach . 1994. "Anaphora Resolution in Machina Translation". Studies in Machine Translation and Natural Language Processing, (Volume 6 "Text and content in Machine Translation: Aspects of discourse representation and discourse

processing") ed. by W. Ramm (ed), 29-52. Luxembourg: Office for Official Publications of the European Community.

81. Pollard, C., and Sag, I.A. Anaphors in English and the Scope of Binding Theory. Linguistics and Philosophy,23, 2 (1992). pp. 261-303.

82. Popowich, F. Reflexives and Tree Unification Grammar. Ph.D. thesis. Universitty of Edinburgh, 1988.

83. Reinhart, T. Anaphora and Semantic Interpretation. Croom Helm Backenham ,Kent. 1983.

84. Rich, Elaine & Susann LuperFoy. 1988. "An architecture for anaphora resolution". Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2), 18-24, Texas, U.S.A.

85. Rich, E., and LuperFoy, S. Anaphora architecture for anaphora resolution. In Proceedings of the Second Conference on Applied NLP (1988).

86. Rico, Celia. Aproximación estadístico-algebraica al problema de la resolución de la anáfora en el discurso. Tesis Doctoral. Universidad de Alicante, 1994.

87. Rico Pérez, Celia. 1994. "Resolución de la anáfora discursiva mediante una estrategia de inspiración vectoral". Proceedings of the SEPLN'94 conference (SEPLN'94). Cordoba, España.

88. Ristad, E.S. The anaphora problem. Departmant of Computer Science. Princeton University, Academic Press, Inc., 1993.

89. Rocha, M. Corpus-based study of anaphora in English and Portuguese. In Corpusbased and computational approaches to discourse anaphora. S.P. Botley and A.M. McEnery Eds., UCL Press. 1997.

90. Rocha, M. Supporting anaphor resolution in dialogues with a corpus-based probabilistic model. In Proceedings of ACL/ EACL workshop on Operational factors in practical, robust anaphor resolution (Madrid, 1997).

91. Sidner, C. Focusing for interpretation of pronouns. American Journal of Computational Linguistics,7 (1981). pp. 217-231.

92. Sidner, C. Focusing in the comprehension of definite anaphora. In Computational Models of Discourse. Brady y Berwick, Eds., MIT. 1983.

93. Sidner, C. L. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Ph.D. thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1979.

94. Sidorov G. Problemas actuales de Lingüística Computacional. *Revista digital universitaria*, UNAM, México. Vol. 2 No. 1, ISSN 1607-6079, marzo 2001.

95. Shalom, L., and Nissim, F. E-Type pronouns, I-sums, and donkey anaphora. Linguistics and Philosophy,17 (1994). pp. 391-428.

96. Strube, M. Processing Complex Sentences in the Centering Framework. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL'96 (Santa Cruz, Ca., 1996).

97. Strube, M., and Hahn, U. Functional centering. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL'96 (Santa Cruz, Ca., 1996).

98. Stuckardt, R. Anaphor Resolution and the Scope of Syntactic Constraints. In Proceedings of the 16th International Conference on Computational Linguistics, COLING'96 (Copenhagen, Denmark, 1996). Vol. II, pp. 937-942.

99. Stuckardt, R. Resolving anaphoric references on deficient syntactic descriptions. In Proceedings of ACL/ EACL workshop on Operational factors in practical, robust anaphor resolution (Madrid, 1997).

100. Suri, L.Z., and McCoy, K.F. RAFT / RAPR and Centering: A comparison and discussion of problems related to processing complex sentences. Computational Linguistics, 20, 2 (1994). pp. 301-317.

101. Sutcliffe, R. A Parallel Distributed Processing Approach to the Representation of Knowledge for Natural Language Understanding. Tesis doctoral. Universidad de Essex, 1989.

102. Walker, M.A., Iida, M., and Cote, S. Japanese Discourse and the Process of Centering. Computational Linguistics,20, 2 (1994). pp. 193-232.

103. Webber, B. L. A Formal Approach to Discourse Anaphora. Ph.D. thesis. Division of Applied Mathematics. Harvard University, Cambridge, MA, 1978.

104. Weizenbaum, J. ELIZA: a computer program for the study of natural language communication between man and machine. Communications of the ACM.9 (1996). pp. 36-45.

105. Williams, S., Harvey, M., and Preston, K. Rule-based reference resolution for unrestricted text using part-of-speed tagging and noun phrase parsing. In Proceedings of the Discourse Anaphora And Resolution Colloquium, DAARC96 (1996).

106. Wilks, Y. A preferential pattern-seeking semantics for natural language inference. Artificial Intelligente,6 (1975). pp. 53-74.

107. Winograd, T. A procedural model of language understanding. In Readings in Natural Language. Grosz et al, Eds., California: Morgan Kaufman.1986.

108. Winograd, T. Understanding Natural Language. Academic Press, Nueva York, 1972.

109. Woods, W. Lunar rocks in natural English: Explorations in natural language question answering. In Linguistic Structures Processing. A. Zampolli, Ed., Nueva York: Elsevier. 1977.

110. Woods, W., Kaplan, M., and Nash-Webber, B. The Lunar Sciences Information System. Cambridge, Mass: Bolt, Beranek y Newman Inc., 1972.

workshop on Operational factors in practical, robust anaphora resolution. 38-45.

10. Baldwin, B., Reynar, J., Collins, M., Eisner, J., Ratnaparkhi, A., Rosenzweig, J., Sarkar, A., and Srinivas. University of Pennsylvania: Description of the University of Pennsylvania system used for MUC-6. In Proceedings of the Sixth Message Understanding Conference (1995).

11. Bobrow. The high-school algebra problem answering system STUDENT with natural language input, 1964.(*)

12. Bobrow, D., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. Gus, a frame driven dialog system. Artificial Intelligence, 8 (1977). pp. 155-173.

13. Botley, S. Comparing demonstrative features in three written English genres. In Proceedings of the Discourse Anaphora And Resolution Colloquium, DAARC96 (1996).

14. Brennan, S., Friedman, M., and Pollard, C. A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, ACL'87 (1987). pp. 155-162.

15. Brown, G., and Yule, G. Discourse Analysis. Cambridge University Press, 1983.

16. Carbonell, James G. & Ralf D. Brown. 1988. "Anaphora resolution: a multi-strategy approach". Proceedings of the 12. International Conference on Computational Linguistics (COLING'88), Vol.I, 96-101, Budapest, Hungria.

17. Carter, D. Control issues in anaphor resolution. Journal of Semantics,7 (1990). pp. 435-454.

8. Carter, D. Interpreting Anaphors in Natural Language. Ellis Horwood, Ed., Chichester, UK, 1987.

9. Charniak, E. Toward a Model of Children's Story Comprehension. Cambridge, Mass: MIT A.I. Lab TR-266. 1972.

20. Chierchia, G. Anaphora and Dynamic Binding. Linguistics and Philosophy,15 (1992). pp. 111-183.

21. Chomsky, N. Knowledge of Language. Praeger, New York. 1986.

22. Chomsky, N. The New Sintaxis: Government and Binding Theory, 1988.

23. Chomsky, N. Lectures on Government and Binding. Foris Publications, Dordrecht, 1981.

24. Collins, M. A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, ACL'96 (1996).

25. Connoly, D., Burger, J., and Day, D. A machine learning approach to anaphoric reference. In Proceedings of the International Conference on New Methods in Language Processing (UMIST, Manchester, 1994).

26. Covington, M. Natural Language Processing for Prolog Programmers. Prentice Hall, 1994.

27. Cullingford, R. E. Inside Computer Understanding: Five Programs Plus Miniatures. Lawrence Erlbaum, Hillsdale, NJ. In R. C. Schank and C. K. Riesbeck, Eds. 1981.

28. Dagan, Ido y Alon Itai (1991). "A statistical filter for resolving pronoun references", Artificial Intelligence and Computer Vision, 125-135.