La desambiguación automática de oraciones pronominales

Ana Fernández

Gloria Vázquez

Irene Castellón

Abstract

El trabajo que presentamos aborda la desambiguación automática de oraciones pronominales en español a partir de la asignación de una representación semántica a cada estructura. A tal fin se ha desarrollado un algoritmo para etiquetar la semántica de las construcciones con el pronombre *se*: reflexivas, recíprocas, anticausativas, medias, etc. Dicho algoritmo se basa en la combinación de diversas fuentes de información lingüística: léxica, morfológica, sintáctica, semántica e información ontológica.

1. Introducción

El marco de nuestro trabajo es el proyecto ReSim[1], que tiene como objetivo principal la construcción de un sistema automático que genere las representaciones semánticas de oraciones de la lengua española. Una de las etapas inciales del proyecto es el estudio de un tipo de oraciones, las construcciones pronominales. Este tipo de estructuras presentan una dificultad adicional al problema de la construcción de las plantillas semánticas, ya que se caracterizan por ser estructuralmente idénticas aunque semánticamente diferentes. En este artículo nos centraremos en la definición de mecanismos para llevar a cabo la desambiguación de forma automática.

La construcción conocida como pronominal se define formalmente como aquella estructura sintáctica en que la forma *se*[2] se combina con un verbo. Desde el punto de vista semántico, este tipo de oraciones pueden expresar una disparidad de significados oracionales, como reflexivo, recíproco, anticausativo, medio, pasivo o impersonal. A veces, la ambigüedad es tal que podría incluso considerarse que determinadas estructuras pronominales son neutras por lo que respecta a determinados significados. Así, en la oración siguiente las interpretaciones anticausativa y pasiva presentan fronteras borrosas. Esta oración (1) podría interpretarse como anticausativa si se entiende que no existe un agente del evento, mientras que también podría interpretarse como pasiva en el caso de que algún agente lo hubiera realizado

(1) Se ha hundido un barco

[1] Representación simbólica de preguntas en lenguaje natural. MCyT (BFF2001-5440-E)

[2] También *me, te, nos* y *os*, pero no *le, lo, la, les, las*. En este artículo nos centraremos en las construcciones con "se" únicamente.

Otras veces, es posible desambiguar por el contexto lingüístico a partir de la información que proveen ciertos elementos, como por ejemplo los adjuntos. En (2) la presencia del SP nos permite interpretar la oración como anticausativa y no como pasiva:

(2) Se han despegado los carteles con la lluvia

Dicha interpretación nos obligará a traducir dicha oración tal como se propone en (3) y no en (4):

(3) All posters came unglued (anticausativa)
(4) All posters were detached (pasiva)

En ocasiones, la ambigüedad es usada de manera consciente por el autor con motivos estilísticos, como en el siguiente ejemplo, donde el propio autor explicita el doble sentido de la oración:

(5) Él o yo estábamos traicionándonos, no sé si el uno al otro o cada uno a sí mismo[3].

Desde el punto de vista del procesamiento del lenguaje natural, y en especial para aplicaciones que requieran una mínima comprensión del lenguaje, como por ejemplo, la traducción automática (TA) o las aplicaciones basadas en diálogos, es necesario dar una representación semántica unívoca. Por este motivo es necesario desambiguar y llegar, en la medida de lo posible, a interpretar semánticamente las oraciones.

Tal y como ya se ha señalado, el objetivo último de nuestro trabajo es la creación de un motor que permita la construcción automática de representaciones semánticas. Para poder llevar a cabo esta tarea es necesario etiquetar las oraciones por lo que respecta a su significado implícito. Es en este marco en el que la asignación de significado al conjunto de la oración se convierte en una tarea esencial. El procedimiento de desambiguación que presentamos se ha desarrollado a partir de un corpus de oraciones pronominales, cuyo estudio nos ha permitido establecer una casuística para desambiguar, de manera automática, los diferentes significados oracionales de estas construcciones.

El presente artículo se estructura del siguiente modo. En primer lugar, estudiaremos brevemente las construcciones pronominales que se tratan en este trabajo (apartado 2) y su relevancia en un corpus del español (apartado 3). En segundo lugar (apartado 4), presentaremos los factores que determinan la heurística. Finalmente (apartado 5), se presentaran unos ejemplos de aplicación de las heurísticas en algunas oraciones del corpus.

2. Las construcciones pronominales

[3] Corpus Lexesp (Lexesp II Acción especial APC96-0125)

En esta sección presentamos los tipos de construcciones pronominales que se han considerado en este trabajo. Todas ellas son relevantes para nuestro fin, ya que desde el punto de vista del significado presentan unas características propias, que las diferencian entre sí. La base de las diferencias semánticas de estas oraciones se fundamenta en la relación que se establece entre el verbo y sus respectivos argumentos y la estructura eventual.

El lenguaje de representación que utilizamos es el propuesto por Jackendoff 1990, las llamadas estructuras léxico-conceptuales (LCS). Este sistema de representación se ha completado en los casos de aquellas construcciones para las cuales no se disponía de una propuesta por parte de este autor. Las estructuras léxico-conceptuales presentan la ventaja de representar adecuadamente la interacción entre sintaxis y semántica y han sido aplicadas para sistemas de procesamiento de lenguaje natural, como la TA (Dorr 1990, Saint-Dizier 1999).

A continuación se describen los diferentes tipos de eventos denotados por las construcciones pronominales que se han considerado.

Eventos reflexivos:

Este tipo de eventos se caracterizan formalmente por la identidad entre dos elementos participantes en el evento, dado que la acción se lleva a cabo por y sobre una misma entidad conceptual, en su totalidad o en una parte de ella. Normalmente, el tipo de entidades que pueden participar en esta construcción pertenecen al tipo semántico humano o animado. Estos eventos admiten la inclusión de la expresión *a sí mismo* para enfatizar la reflexividad. En este caso, el pronombre puede estar flexionado en las diferentes *personas (me, te, se, nos, os*). Veamos algunos ejemplos:

(6) a. Juan se viste (a sí mismo)
b. Ana se peina (a sí misma)
b. Yo me pregunto (a mí mismo) qué ocurrirá

Para su representación semántica hemos escogido la representación propuesta en Jackendoff, en la que la el sujeto y el objeto (pronombre) se representan como la misma categoría conceptual mediante la utilización de α. A continuación se presenta la LCS para (6a):

$$[_{event}\ \text{CAUSE}\ ([\ ]^{\alpha}_{i},[\text{GO}([\alpha]_{j},[\text{TO}[\text{IN}[\text{CLOTHING}]]_{k}])])]$$

Eventos recíprocos:

Un evento recíproco describe una acción múltiple, es decir, es un evento complejo que implica la realización de varios eventos y se caracteriza porque los participantes implicados presentan los papeles invertidos. Para que un evento sea interpretado recíprocamente es necesario que el verbo esté en plural y es habitual la aparición de las expresiones *mutuamente* o *el uno al otro*. El pronombre admite también diferentes formas personales (*nos, os, se*).

(7) a. Pedro y Juan se golpearon
b. Pedro y Ana se acarician (el uno al otro)
c. Nunca se saludan por la calle

Para representar este tipo de construcción hemos reelaborado la LCS utilizada en la interpretación reflexiva para poder expresar este tipo de significado. En el siguiente ejemplo, se presenta la representación semántica de (7a):

$$[_{event}\ [_{event}\ \text{CAUSE}\ ([\ ]^{\alpha}_{i},[\text{INCH}[\text{BE}\ ([\text{STICK}],[\text{AT}\ [\beta]_{j}]])]])]$$
$$\&\ [_{event}\ \text{CAUSE}\ ([\ ]^{\beta}_{i},[\text{INCH}[\text{BE}\ ([\text{STICK}],[\text{AT}\ [\alpha]_{j}]])]])]]$$

Esta conceptualización implica la existencia de dos eventos en cada uno de los cuales uno de los actores es sujeto y el otro objeto, con los papeles intercambiados en cada uno de los eventos. Tal y como sucedía en los eventos reflexivos, en los eventos recíprocos los participantes son también usualmente de tipo animado. Cabe señalar que en muchas ocasiones se da una cierta ambigüedad entre las interpretaciones reflexiva y recíproca, siendo más habitual la primera de éstas.

Eventos anticausativos:

Eventos anticausativos son aquellos en los que se pretende enfatizar el cambio o la modificación que ha sufrido una entidad:

(8) a. El pantano se llenó de agua
b. Me he asustado
c. La pintura se ha oscurecido

Como puede verse en estas oraciones, la causa que desencadena el cambio o modificación de la entidad puede quedar inexpresada. Los motivos por los que se elige esta construcción pueden ser diversos: o bien porque se desconoce la causa que provoca el cambio, o bien porque ésta es conocida de forma generalizada por parte de los participantes en la comunicación y, por lo tanto, se considera no relevante.

Para este tipo de oraciones la representación semántica elegida es la que presentamos a continuación para la oración (8a):

$$[_{event}\ \text{INCH}\ [_{state}\ \text{BE}\ ([\ \ ]_{k}\ [\text{IN}\ [\ \ ]_{j}])]]$$

Esta representación corresponde al subevento que expresa el cambio de estado que está implícito en la representación de un evento causativo, el cual actúa como desencadenante del anterior. Según la representación propuesta por Jackendoff, el *pantano* representa el locativo, por lo tanto corresponde a j y el elemento que se halla en este lugar es el *agua*, *k*. Por lo que se refiere al pronombre, admite las mismas formas que la construcción reflexiva.

Eventos antiagentivos:

Como en el caso de las construcciones anticausativas, se trata de eventos en los que el foco informativo recae sobre la entidad a la que va dirigida la acción. Ahora bien, a diferencia del caso anterior, en los eventos antiagentivos pronominales, por un lado, el elemento que provoca la acción no es causativo y, por otro lado, no es necesario que se modifiquen las propiedades características de la entidad sobre la que se predica. Desde el punto de vista formal, la antiagentiva se distingue de las anteriores por el hecho de que siempre presenta la forma pronominal *se*.

Entre los eventos antiagentivos incluimos las estructuras pasivas y las impersonales (Moreno Cabrera 1991). La diferencia entre estas dos construcciones es puramente formal: mientras que las primeras tienen un sujeto sintáctico, las segundas no, ya sea porque el objeto presenta la forma de SP (10a y 10b), porque el verbo puede ser usado intransitivamente (10c) o, menos habitualmente, porque no concuerden el verbo y el objeto, representado por un SN.

- Pasivas:

(9) a. Creemos que los cadaveres se hundieron con la ayuda de un peso
b. Los sobres se han repartido por la mañana
c. Se ha trasladado a los heridos al hospital más cercano

- Impersonales:

(10) a. Creemos que se hundió a las víctimas con la ayuda de un peso
b. Se ha trasladado a los heridos al hospital más cercano
c. Se pagó en efectivo

La representación semántica del segmento *los cadáveres se hundiero*n de (9a) y *se hundió a las víctimas* de (10a) es la misma[4]:

$[_{event}$ CAUSE ([ ],[GO([$]_{i/j}$,[DOWN FROM SURFACE OF WATER])])]

Estativa pronominal:

El tipo de estados que se describen en esta sección se denominan también estados de eventos potenciales (Croft 1997). Se entiende como estados de eventos potenciales aquellas situaciones denotadas por verbos que son habitualmente eventivos pero que pueden expresar un estado cuando las coordenadas espacio-temporales no están especificadas. Algunos verbos en español requieren construcciones pronominales para expresar este tipo de estados:

[4] Excepto en la correlación entre los constituyentes semánticos y su realización sintáctica, ya que en (9a) 'los cadáveres' actúa como sujeto y en (10a) 'las víctimas' es el objeto.

(11) a. La madera se estropea con la humedad / con facilidad
b. Esta fruta no se come
c. Los jueves se come paella

Dentro de este tipo de construcciones hemos considerado diferentes construcciones que tienen en común su carácter estativo, como son, por ejemplo, la media (11a y 11b) y la habitual (11c). Todas ellas tienen en común la expresión de una propiedad de una entidad. El carácter estativo de estas construcciones viene enfatizado por el uso de un tiempo no marcado (típicamente, el presente) y por la aparición de determinados adjuntos. Estos adjuntos pueden expresar la manera en que la propiedad se modifica (*con facilidad*, 11a), el instrumento o causa (*con la humedad*, 11a) o la frecuencia (*los jueves*, 11b). El carácter estativo puede ser conferido también por la presencia de un verbo modal (*poder*) o la negación (como en 11b).

Proponemos la siguiente representación para este tipo de oraciones:

$$[_{\text{state}}\ \text{BE}\ ([_{\text{thing}}\ \ ]_j,([\text{AT}\ [_{\text{eventive\_property}}\ \ ]], [_{\text{manner/time/...}}\ ]))]$$

3. Las oraciones pronominales en un corpus

Se ha llevado a cabo el estudio sobre las frecuencias de las oraciones pronominales a partir de una colección de textos obtenidos de Internet y que pertenecen fundamentalmente al ámbito periodístico. Se ha constatado una alta frecuencia de aparición de patrones pronominales en castellano, ya que esta lengua utiliza la pronominalización para expresar variedad de significados oracionales, que en otras lenguas se expresan mediante diferentes mecanismos gramaticales y morfológicos.

El corpus está compuesto de un total 453.232 oraciones, de las cuales se han recogido un total de 48.462 construcciones que contienen pronombres y de éstas 35.886 son las que contienen *me*, *te*, *se*, *nos* y *os* (aproximadamente, un 7% del corpus). El conjunto de oraciones pronominales se ha refinado con la selección únicamente de aquellos patrones en los que el pronombre tiene la forma *se*, con lo cual el corpus con el que hemos trabajado está constituido por 32.465 oraciones.

Este corpus se ha dividido en dos: una parte se ha destinado al estudio de las oraciones y la otra a la evaluación. El primero nos ha servido para realizar el análisis que nos ha permitido adquirir el conocimiento necesario para establecer la heurística. El corpus de evaluación se ha utilizado por ahora para aplicar la heurística con el objetivo de refinar y mejorar las reglas. En el futuro se prevé utilizar este corpus para aplicar automáticamente las heurísticas y evaluar su calidad.

4. Algoritmo propuesto para la resolución de la desambiguación de pronominales

En este apartado vamos a presentar la heurística, las fuentes de conocimiento que se requieren para su correcta aplicación, el tipo de reglas que se han realizado y algunos ejemplos de desambiguación de oraciones en corpus.

El procedimiento que presentamos tiene como objetivo final la asociación de una categoría semántica oracional (causativa, antiagentiva, etc) a una oración y constituye, por lo tanto, un sistema de desambiguación oracional. La información que se utiliza en este procedimiento es de dos tipos. Por un lado, se utilizan fuentes de conocimiento morfológico, léxico y ontológico. Por otro lado, el sistema cuenta con un conjunto de heurísticas, que constituyen el elemento central del mismo, ya que es en este módulo en el que se lleva a cabo propiamente la tarea de desambiguación. Durante dicho proceso las heurísticas utilizan las diferentes fuentes y es mediante la combinación de todo este conocimiento que se asignan, eliminan o priorizan diferentes interpretaciones.

4.1 Fuentes de conocimiento

Las diferentes fuentes de conocimiento, tal y como se ha mencionado anteriormente, aportan información de diferente tipo. La información sintáctico-semántica está codificada en una base de datos verbal. Para poder comparar la oración de entrada con los datos de subcategorización almacenados en dicha base es preciso un tratamiento morfo-sintáctico de la oración. Por otro lado, en lo que refiere a las restricciones de selección impuestas por determinados predicados o construcciones, el proceso precisa de una clasificación léxico-ontológica.

En la base de datos verbal (Vázquez et al. 2001, Fernández et al. 2002) desarrollada dentro del proyecto Volem[5], la unidad básica es el sentido verbal, entendido desde una perspectiva sintáctico-semántica. Desde esta perspectiva los criterios para el establecimiento de sentidos producen menos grado de ambigüedad que en otros recursos, como EuroWordNet (EWN) (Vossen 1999). Cada sentido verbal se define mediante la clase semántica a la que pertenece (Vázquez et al. 2000) y su estructura argumental, expresada mediante papeles temáticos y el conjunto de esquemas sintáctico-semánticos que determinan la interpretación de la oración. Cada uno de estos esquemas incluye la lista categorial de los constituyentes asociada a la etiqueta semántica que representa el significado oracional (anticausativa, impersonal, pasiva,etc.).

Las ventajas del uso de este recurso durante el proceso de desambiguación son diversas. En primer lugar, el hecho de disponer de esquemas de subcategorización especificados permite saber *a priori* qué estructuras son plausibles y, por lo tanto, reduce el número posible de interpretaciones que se toma como punto de partida. En segundo lugar, la integración de información sintáctica y semántica en el lexicón permite resolver parte de la ambigüedad oracional a partir de marcas formales.

Otra fuente de conocimiento utilizada es el resultado de la aplicación de dos procesos de análisis: el morfológico y el sintáctico, que agrupa en constituyentes básicos las unidades de la oración (Atserias et al. 1998). Esta anotación del texto morfo-sintáctica se utiliza a lo

[5] ABM/acs/XTI-CTP 2000-1.

largo del proceso de desambiguación para comprobar algunas categorías, lemas o secuencias de ellos, condicionando de este modo la aplicación de algunas heurísticas.

Otro de los recursos del sistema es EWN, que se utiliza como ontología para obtener conocimiento del mundo, especialmente a partir del campo semántico y de las relaciones de hiponimia. Este recurso servirá para crear bolsas de palabras para expresar preferencias selectivas y tipos de objetos del mundo implicados en los eventos con el fin de establecer condiciones en la asignación de esquemas. La consulta de este tipo de información permite definir de forma más precisa el procedimiento de desambiguación y, de este modo, alcanzar mejores resultados en el reconocimiento de la semántica de la frase.

Por último, utilizamos un segmentador de texto (Alonso y Castellón 2001) que permite delimitar y etiquetar las partes del mismo utilizando marcadores del discurso (signos de puntuación, conjunciones, locuciones preposicionales, adverbiales, etc). Estos segmentos no son nunca superiores a una oración (segmentos entre dos puntos) y pueden ser anidados o independientes. Los segmentos corresponden básicamente a oraciones subordinadas y oraciones principales, aposiciones, grupos adverbiales, preposicionales, etc. A partir de los resultados de esta herramienta, se realiza una segunda selección de los segmentos que contienen ocurrencias de 'se' sobre los cuales se aplicará la heurística que presentamos a continuación.

4.2 Heurística

La heurística se compone de reglas que, aplicadas sobre las oraciones, proponen una interpretación de las mismas. Para poder llevar a cabo esta tarea, el primer paso consiste en realizar una consulta a la base de datos léxica con el fin de obtener la información que puede ser relevante. El resultado es una la lista con las posibles interpretaciones oracionales para cada verbo. Además, se obtiene también otro tipo de información semántica (restricciones selectivas, papeles temáticos y preposiciones), que puede ser utilizada en otras fases del proceso.

Las reglas que configuran la heurística pueden clasificarse de diversas formas: según la acción que realizan, la información que manejan o la construcción sobre la que actúan. Por lo que se refiere al tipo de acción, existen dos tipos de reglas: de restricciones y de preferencias. Las primeras pueden eliminar o asignar interpretaciones definitivas. Las segundas ponderan positiva o negativamente determinadas interpretaciones.

El proceso finaliza en uno de los siguientes supuestos:

- una regla de restricción asigna la interpretación definitivamente
- se eliminan todas las interpretaciones menos una
- se ordenan las interpretaciones según la ponderación
- no se llega a proponer ninguna interpretación como definitiva ni preferible.

Como hemos dicho, el conocimiento que utilizan estas heurísticas es siempre de tipo lingüístico –extraído del análisis morfo-sintáctico de la oración y de la base de datos verbal– y ontológico –principalmente, clases de palabras derivadas de EWN. A partir de esta información, se utilizarán condiciones morfológicas sobre determinadas categorías y subcategorías, se comprobarán determinadas condiciones sintácticas como la concordancia, la presencia de determinados sintagmas o palabras, etc.

Las heurísticas desarrolladas tienen un doble objetivo. En primer lugar, se persigue determinar la interpretación semántica básica de la construcción pronominal entre las cinco posibles tenidas en cuenta: *reflexivas, recíprocas, anticausativas, pasivas* e *impersonales.* En la actualidad se está trabajando en la ampliación de heurísticas para etiquetar otras oraciones pronominales, como las *procesuales* o las de *dativo de interés*.

En segundo lugar, otro grupo de heurísticas determinará el aspecto de la construcción con el fin de diferenciar las construcciones *estativas* de las *eventivas*. Con ello se pretende dar cuenta, por ejemplo, de la llamada construcción *media*.

La heurística diseñada está formada por un total de 22 reglas. La aplicación de estas reglas se ha estructurado en cinco fases, que responden al orden del proceso de aplicación y a la acción de la heurística: consulta a la base de datos (fase 1), aplicación de las condiciones (en la fase 2 eliminan, en la 3 determinan y en la 4 ponderan) y aplicación de heurísticas aspectuales (fase 5).

En la **primera fase** se pretende limitar de entrada las expectativas en función de los requisitos léxicos de cada verbo. El objetivo es crear una lista de posibles interpretaciones de la oración (para cada sentido verbal) a partir de la consulta a la base de datos verbal con el fin de obtener el conjunto de posibles oraciones pronominales para ese lema según el sentido y su estructura temática.

En algunos casos, si el verbo no presenta ambigüedad y sólo admite una posible construcción pronominal con una única interpretación, se obtiene ya un resultado. En el caso de que el verbo tenga diferentes sentidos, puede ocurrir que se elimine alguno si en la entrada correspondiente no está codificada ninguna construcción pronominal como plausible. De este modo, la información sintáctico-semántica estaría contribuyendo al campo de la desambiguación léxica (Word Sense Disambiguation). Por ejemplo para la frase siguiente (12), el sistema descartaría los sentidos codificados en Volem como *llenar563 (el trabajo la llena)* y *llenar279 (el primer plato le ha llenado)*, siguiendo el proceso con *llenar900 (María llena la furgoneta de paquetes)* y *llenar925 (la gente llenó la plaza)*, que sí admiten la construcción pronominal.

(12) La sala se llenó con los acordes de una sinfonía

La **segunda fase** consiste en la eliminación de determinadas interpretaciones. Para ello, se aplican reglas que nos permiten eliminar candidatos de la lista, es decir, reducir las posibilidades. Concretamente, estas reducciones se aplican para suprimir las interpretaciones recíprocas, reflexivas e impersonales. Si en la aplicación de estas reglas se redujese la lista a un solo candidato, se daría por terminado el proceso.

En comparación con el resto, estas tres construcciones presentan unas marcas formales que las caracterizan inequívocamente. Así pues, la ausencia de dichas marcas o la presencia de marcas opuestas a las requeridas para la interpretación de dichas construcciones nos permite rechazarlas como candidatas a ser la interpretación de la oración. Una de las condiciones que se utiliza en esta fase son las características requeridas para algunos argumentos en algunas construcciones. Por ejemplo, en el caso de oraciones reflexivas y recíprocas es necesario un SN preverbal humano (o animado). De esta manera, oraciones como (13a) y (13b) se diferencian porque a (13a) no se le podrá asignar una interpretación reflexiva mientras que la oración (3b) sí que la permitirá:

(13) a. El chicle se ha enganchado en el suelo
b. María se ha enganchado una pegatina

Otro tipo de información semántica que se utiliza para refinar el juicio sobre las construcciones reflexivas y recíprocas es la clase verbal. Así, por ejemplo, al ceñirnos a los verbos de trayectoria (Vázquez et al. 2000), siempre que se detecte la presencia de *se* seguido de otro pronombre, descartamos una de estas dos interpretaciones para las siguientes oraciones:

(14) a. Se me dijo tarde la verdad
b. No se les envió la documentación a tiempo

Otro tipo de construcción que puede ser descartada a partir de marcas formales es la construcción impersonal. Por ejemplo, cuando en la construcción de un verbo obligatoriamente transitivo se detecta un SN que concuerda con el verbo y éste no es temporal se descarta esta construcción. Esta heurística se aplicaría en la siguiente oración:

(15) Se han enviado las cartas

La **fase tercera** se compone de heurísticas que permiten deducir la interpretación de la construcción pronominal y, por lo tanto, determinan un resultado seguro, es decir, la aplicación de estas reglas implica el fin del proceso. Estas heurísticas se basan en la combinación de la presencia de determinadas marcas léxicas.

Por ejemplo, la heurística que determina la interpretación recíproca comprueba que existan determinados elementos en el contexto, como puede ser *mutuamente* o *entre ellos*.

(16) Se felicitaron mutuamente

En el caso de las construcciones agentivas con pronombre ‘se’, una manera para determinar este tipo de interpretación es a partir de la falta de concordancia entre este este pronombre y el verbo en los casos en que este no está en tercera persona:

(17) Se lo desabroché

La **fase cuarta** consiste en la aplicación de heurísticas que ponderan determinadas construcciones en función de determinadas características. En este caso el sistema creará una lista con las interpretaciones ordenadas según esta ponderación. Hasta el momento hemos utilizado ponderaciones positivas, es decir, que asignan un peso a algunas interpretaciones en función de ciertas características del verbo o de la oración. El proceso se da por finalizado aunque no haya una única interpretación definitiva.

La ponderación puede ser más o menos fuerte, en función del grado de rigidez de la restricción. Así, la interpretación anticausativa se pondera positivamente (2 puntos) cuando el SN es de tipo abstracto, como en (18). Un caso de ponderación fuerte (4 puntos) es el de la reflexiva en el caso de que se cumpla el esquema "*se* V SN" y el núcleo del SN es una parte del cuerpo o una prenda de vestir, como en (19):

(18) Un pequeño problema se puede convertir en un verdadero conflicto

(19) Se ha pintado las uñas

La **quinta fase**, en la que se etiquetan las construcciones pronominales desde el punto de vista eventivo, está en desarrollo. En este momento el resultado de estas heurísticas es en muchas ocasiones determinante pero en otras, las estativas, decide de forma ponderativa. En general, se tiene en cuenta el tiempo verbal según sea marcado (evento) o no marcado (estado). Además, también se utilizan algunas marcas léxicas, como la presencia de determinados adverbios. Por ejemplo, un tiempo presente y el adverbio *no* se utilizan como criterios para priorizar la interpretación estativa:

(20) Esta fruta no se come

5. Aplicación de la heurística

A continuación presentamos un ejemplo detallado del funcionamiento de la heurística, siguiendo el orden establecido en las diferentes fases, a través de una oración extraída del corpus:

(21) 'La euforia se desató en varios idiomas a las doce de la noche.'

El primer paso consiste en el análisis morfológico y sintáctico (superficial) de la oración. La falta de un análisis sintáctico completo de la oración nos obliga a establecer algún tipo de límite contextual para poder aplicar condiciones sobre los elementos que acompañan al verbo. Por ello todo el procedimiento se realiza sobre segmentos textuales que contienen una unidad verbal en forma pronominal. En el ejemplo que se presenta el segmento coincide con la oración, pero en otros casos se trata de segmentos de orden inferior.

[La_la_TDFS0 euforia_euforia_NCFS000]sn
[se_él_PP3CN000 desató_desatar_VMIS3S0]gv
[en_en_SPS00 varios_varios_DI3MP00 idiomas_idioma_NCMP000]sp
[a_a_SPS00 las_la_TDFP0 doce_doce_MCCP00 de_de_SPS00 la_la_TDFS0 noche_noche_NCFS000]sp

A partir de esta información se extrae el verbo principal de la oración para proceder a la consulta del lema correspondiente en la base de datos. De la entrada verbal se seleccionan aquellas estructuras pronominales en las que puede participar aquel verbo y se crea una lista. Esta lista se complementa con información correspondiente a otras características que pueden ser de interés, como las restricciones selectivas, las preposiciones y los papeles temáticos. Este procedimiento se realiza para todos los sentidos asociados al lema verbal. El lema *desatar* está codificado con dos sentidos en dicha base (*desatar113* y *desatar918*), representados en las figuras 1 y 2, respectivamente. En nuestro ejemplo (21) el sentido se corresponde con *desatar918* y la interpretación correcta es la anticausativa.

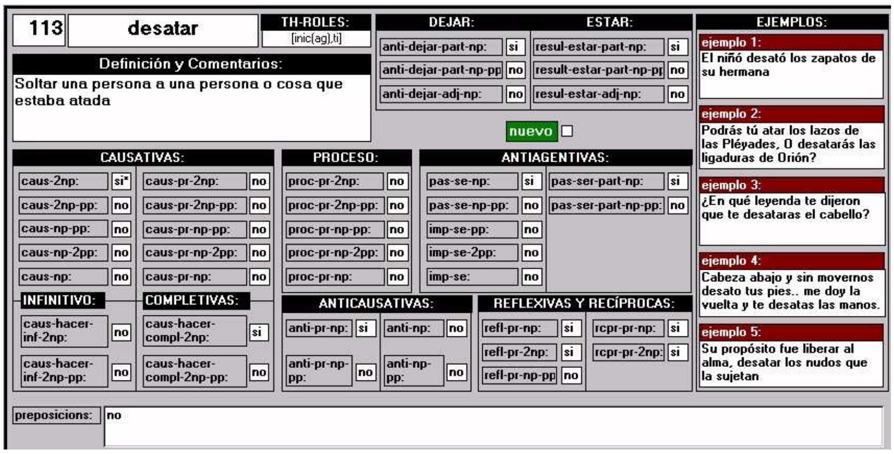

Fig. 1 Entrada léxica 113: desatar

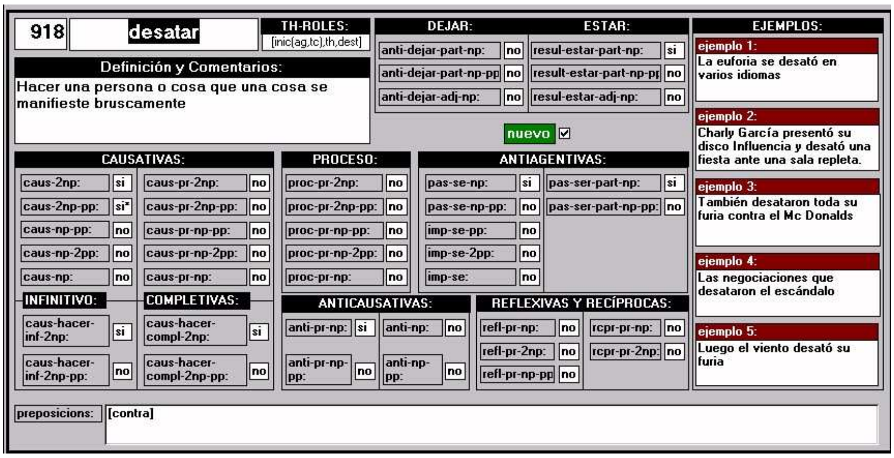

Fig. 2 Entrada léxica 918: desatar

A continuación se presentan las listas de construcciones pronominales para cada sentido:

DESATAR 113		**DESATAR 918**	
refl-pr-np	reflexiva	anti-pr-np	anticausativa
refl-pr-2np	reflexiva	pas-se-np	antiagentiva pasiva

rcpr-pr-np recíproca rcpr-pr-2np recíproca anti-pr-np anticausativa pas-se-np antiagentiva pasiva	
Prep: [ ]	Prep: [contra]
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

A partir de aquí se aplican las condiciones. Primero, en la fase 2 se eliminan las interpretaciones reflexiva, recíproca e impersonal, si fuera el caso. En el ejemplo que nos ocupa al aplicar la regla referente al tipo semántico del SN se eliminarán estas construcciones en la lista correspondiente a la entrada 113, con lo que las quedarían modificadas de la siguiente forma:

DESATAR 113	**DESATAR 918**
~~refl-pr-np reflexiva~~ ~~refl-pr-2np reflexiva~~ ~~rcpr-pr-np recíproca~~ ~~rcpr-pr-2np recíproca~~ anti-pr-np anticausativa pas-se-np antiagentiva pasiva	anti-pr-np anticausativa pas-se-np antiagentiva pasiva
No	PREP: contra
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

La fase 3, es decir , el conjunto de reglas que determinan, no se aplicará dado que la oración no cumple ninguna de las condiciones. A continuación se aplican las reglas que ponderan determinadas interpretaciones (fase 4). En este caso estas reglas son aquellas referentes a las restricciones de selección que ponderan la interpretación anticausativa, ya que el tipo semántico del SN es abstracto. Así, las posibilidades quedan alteradas del modo siguiente:

DESATAR 113	**DESATAR 918**
anti-pr-np anticausativa **+2** pas-se-np antiagentiva pasiva	anti-pr-np anticausativa **+2** pas-se-np antiagentiva pasiva
Prep: [ ]	Prep: [contra]
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

Otra de las reglas de ponderación es la que tiene en cuenta los papeles asignados en las entradas verbales. Por un lado, cuando un verbo admite la anticausativa pero tiene codificado como rol del argumento el agente y no la causa, se prioriza la interpretación anticausativa, como ocurre en el caso de *desatar113*:

DESATAR 113	**DESATAR 918**
anti-pr-np anticausativa **+6** pas-se-np antiagentiva pasiva	anti-pr-np anticausativa +2 pas-se-np antiagentiva pasiva

Prep: [ ]	Prep: [contra]
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

En cambio, las construcciones pasivas se ponderarán en aquellos casos en los que el orden declarado de los papeles corresponda al agente en primer lugar, como ocurre en *desatar918*:

DESATAR 113	**DESATAR 918**
anti-pr-np anticausativa +6 pas-se-np antiagentiva pasiva	anti-pr-np anticausativa +2 pas-se-np antiagentiva pasiva **+1**
Prep: [ ]	Prep: [contra]
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

El paso siguiente es eliminar de la lista aquellas interpretaciones menos puntuadas:

DESATAR 113	**DESATAR 918**
anti-pr-np anticausativa +6 ~~pas-se-np antiagentiva pasiva~~	anti-pr-np anticausativa +2 ~~pas-se-np antiagentiva pasiva) **+1**~~
Prep: [ ]	Prep: [contra]
Roles: (inic(ag),ti)	Roles: (inic(ag,tc),th dest)

Finalmente se aplican las reglas de la fase 5 que es la que nos dan información sobre el tipo eventivo. En este caso se descarta una interpretación estativa ya que el tiempo es marcado, proporcionando el siguiente resultado:

Construcción anticausativa eventiva.
Oración: "La euforia se desató en varios idiomas a las doce de la noche."
Verbo: desatar 918/ desatar 113

En este ejemplo hemos visto que nuestro sistema no necesariamente llega a una desambiguación del sentido verbal. Se determina una interpretación anticausativa pero no de qué verbo. Se prevé incorporar en la base información relativa a las preferencias selectivas. En el ejemplo presentado, dicha información hubiera sido crucial para la desambiguación del sentido en uso, ya que la *desatar113* requiere un SN-tema de tipo físico, mientras que en *desatar918* el tipo de este constituyente sería abstracto, como ocurre en el ejemplo (*euforia*).

6. Conclusiones y líneas futuras

En este artículo se ha presentado un algoritmo para la desambiguación de estructuras pronominales del español, que forma parte de una investigación más amplia sobre la desambiguación oracional de esta lengua. Dicho algoritmo, combinando diferentes fuentes de conocimiento, consigue realizar la desambiguación de forma eficiente. Hay que tener en cuenta que en muchas ocasiones la ambigüedad es irresoluble y es parte de la comunicación humana. En otras ocasiones, la mejora del sistema requeriría la selección de un contexto más amplio para el análisis.

El campo de la desambiguación oracional ha sido muy explorado respecto a la estructura y la jerarquía de constituyentes, pero no en relación con la semántica de la construcción. Consideramos que esta aproximación puede ser útil para diferentes aplicaciones actuales de PLN y constituye un valor añadido especialmente para los sistemas de traducción automática.

Creemos que, de todas las fuentes utilizadas, la información sintáctico-semántica proporcionada por la base de datos verbal es crucial en la tarea de desambiguación, aunque actualmente estamos realizando estudios para comprobar hasta qué punto las diferentes fuentes de información participan en la desambiguación. Para ello tenemos previsto incorporar en la implementación la posibilidad de activar o desactivar las diferentes fuentes de conocimiento y poder de esta manera evaluar el beneficio real que aporta cada una de ellas.

En general, se prevé continuar el trabajo aquí expuesto realizando experimentos masivos de desambiguación y no descartamos que las heurísticas presentadas sean ampliadas o modificadas en vista de los resultados obtenidos.

Bibliografía

- Alonso, L. e I. Castellón, (2001). "Towards a delimitation of discursive segment for Natural Language Processing applications". *International Workshop on Semantics, Pragmatics and Rhetorics.*

- Atserias, J., J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M. A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, J. Turmo (1998) "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text". *First International Conference on Language Resources and Evaluation* (LREC'98).

- Croft, W. (1997) "Possible verbs and the structure of events". En S. L. Tsohatzidis, (ed.). *Meaning and Prototypes*. Nueva York: Routledge, pp. 49-73.

- Dorr, B. J.(1990) *Lexical Conceptual Structure and Machine Translation*. Ph.D., MIT Internal Report, Department of Electrical Engineering and Computer Science, Cambridge, MA.

- Jackendoff , R. (1990) *Semantic Structures*. Cambridge, MA: MIT Press.

- Moreno Cabrera, J. C. (1991*) Curso universitario de lingüística general. Tomo I: Teoría de la gramática y sintaxis general.* Madrid: Síntesis.

- Fernández, A., P. Saint-Dizier, G. Vázquez, F. Benamara y M. Kamel (2002), "The VOLEM Project: a Framework for the Construction of Advanced Multilingual Lexicons". *Proceedings of the Language Engineering Conference* Hyderabad, India.

- Saint-Dizier, P. (1999) "An introduction to the lexical semantics of predicative forms". En P. Saint-Dizier, ed., *Predicative Forms in Natural Languages and in Lexical Knowledge Bases*. Holanda: Kluwer, pp. 1-52.

- Vázquez, G., A. Fernández y M. A. Martí (2000) *Clasificación verbal. Alternancias de diátesis*. Quaderns de Sintagma, Universitat de Lleida.

- Vázquez, G., A. Fernández y M. A. Martí (2001) "Formalización de un modelo diatético", *El verbo: entre el léxico y la gramática*. Lugo: Tris-Tram, pp. 205-219.

- Vossen, P (ed.) (1999) "EuroWordNet General Document. EuroWordNet" (LE2-4003, LE4-8328), Part A, Final *Document, EWN* D032D033/2D014.