Computación y Sistemas Vol. 6 No.2 pp. 143 - 152.
© 2002, CIC - IPN. ISSN 1405-5546 Impreso en México

PH. D. THESIS ABSTRACT

Análisis Sintáctico Conducido por un Diccionario de Patrones de Manejo Sintáctico para Lenguaje Español

A Syntactic Analyzer Driven by a Government Patterns Dictionary for Spanish Language

Graduated: Sofía Natalia Galicia Haro
Graduated on November 29, 2000
Centro de Investigación en Computación, IPN
Avda. Juan de Dios Bátiz s/n esq. Miguel Othón de Mendizabal
Unidad Profesional Adolfo López Mateos C.P. 07738
Del. Gustavo A. Madero, México, D.F.
e-mail: sofia@cic.ipn.mx
Advisor 1: Alexander F. Gelbukh
Centro de Investigación en Computación-IPN, México
e-mail: gelbukh@cic.ipn.mx
Advisor 2: Igor A. Bolshakov
Centro de Investigación en Computación-IPN, México
e-mail: igor@cic.ipn.mx

Resumen

En el análisis sintáctico de textos, mediante computadora, el problema más importante a resolver es la ambigüedad estructural. Con los métodos actuales se obtiene una gran cantidad de variantes cuando se analizan textos sin restricciones. Para eliminar las estructuras incorrectas proponemos un modelo de análisis sintáctico y desambiguación que considera un esquema basado en tres diferentes fuentes de conocimiento del lenguaje, de las cuales la fuente principal dirige el análisis mediante conocimiento lingüístico descrito en un diccionario de patrones de palabras del español, principalmente verbos aunque también se consideran algunos adjetivos y sustantivos. En este artículo también presentamos las características del español consideradas para describir los patrones de palabras y la nueva descripción con información cualitativa, así como el algoritmo desarrollado para la adquisición de información lingüística a partir de un corpus de textos y los resultados obtenidos al aplicar esta información en el análisis de un conjunto de oraciones.

Palabras clave: Análisis Sintáctico, Desambiguación Sintáctica, Patrones de Manejo Sintáctico, Valencias Sintácticas.

Abstract

The most important problem to resolve in natural language parsing by computational means is structure ambiguity. A great quantity of variants is obtained in syntactic analysis of unrestricted texts with current methods. In order to eliminate the erroneous variants we propose a syntactic analysis and disambiguation model which considers a scheme based on three different language knowledge sources. The main source drives the parsing by linguistic knowledge described in a Spanish government patterns dictionary, mainly for verbs but some adjectives and nouns are also considered. In this article we also present the characteristics of Spanish to describe that patterns, the new description with qualitative information, the developed algorithm to acquire linguistic knowledge from a text corpus and the results obtained when applying this information to parse a collection of sentences.

Keywords: Syntactic Analysis, Syntactic Disambiguation, Government Patterns, Syntactic Valences.

1 Introducción

En el análisis sintáctico la tarea principal es describir cómo las palabras de la oración se relacionan y cuál es la función que cada palabra realiza en esa oración, es decir, construir la estructura de la oración de un lenguaje. Las frases posibles de un lenguaje natural son secuencias gramaticales, es decir, que obedecen leyes gramaticales, sin conocimiento del mundo. Establecer métodos que determinen únicamente las secuencias gramaticales en el procesamiento lingüístico de textos por computadora ha sido el objetivo de los formalismos gramaticales. Se han considerado dos enfoques para describir formalmente la gramaticalidad de las oraciones: las dependencias y los constituyentes.

En la obtención de la estructura de las oraciones mediante estos formalismos, las combinaciones de los distintos complementos en la oración presentan cierta complejidad. Por ejemplo, en la frase *Compró el niño un libro en diez pesos en la tienda XX a un lado del metro Juárez a un vendedor alto de mal humor*, existen seis grupos preposicionales (*en la tienda*, *del metro Juárez*, etc.) introducidos con solo tres preposiciones, *a*, *en*, *de*, y dos grupos nominales (*el niño*, *un libro*). Las posibles combinaciones no son aleatorias pero estos complementos o grupos lingüísticos pueden ir enlazados en diferentes combinaciones, unidos al verbo o a sustantivos de los diferentes grupos de la oración, por ejemplo: *Compró un libro*, *Compró en diez pesos*, *Compró en la tienda XX*, *Compró a un vendedor alto*, *la tienda XX a un lado del metro Juárez*. Mientras para un hablante nativo es obvio cómo se relacionan los complementos, para una computadora son posibles todas las variantes: *Compró a un lado*, *Compró del metro Juárez*, *Compró de mal humor*, *el niño en la tienda XX*, etc.

Para eliminar las estructuras incorrectas proponemos un modelo de desambiguación sintáctica basado en tres diferentes fuentes de conocimiento del lenguaje En este trabajo, presentamos las características del español para

describir los patrones de palabras, principalmente verbos, que dirigirán el análisis sintáctico. Enseguida presentamos el modelo y por último el algoritmo desarrollado para la adquisición de la información lingüística de los patrones a partir de un corpus de textos y los resultados obtenidos.

2 Descripción de Valencias

2.1 Enfoques de Descripción

Los constituyentes es el enfoque donde las oraciones se analizan mediante un proceso de segmentación y clasificación. Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso sucesivamente hasta obtener las partes de la palabra indivisibles dentro de la gramática. Por ejemplo, la frase *los niños pequeños estudian pocas horas* se divide en el grupo nominal *los niños pequeños* más el grupo verbal *estudian pocas horas*, este último a su vez, se divide en el verbo *estudian* más el grupo nominal *pocas horas* y así sucesivamente.

Bajo este enfoque, aunque existe un número finito de palabras en el lenguaje, es posible generar un número infinito de oraciones mediante reglas, que también se emplean para la comprensión del lenguaje. Sin embargo, se generan mucho más secuencias de palabras de las que realmente quieren producirse. Por ejemplo, una regla para definir grupos nominales en el español es: un artículo indefinido, seguido de un sustantivo y a continuación un grupo preposicional. Sin embargo, esta regla define tanto *la plática sobre la libre empresa* como **la solidaridad sobre la libre empresa* siendo ésta última una secuencia no gramatical.

En este enfoque, una información importante para el análisis sintáctico es la definida como subcategorización, referida a los complementos que una palabra rectora puede tener y la categoría gramatical de ellos. Esta información se ha agrupado en patrones que describen la composición de los complementos posibles para diferentes verbos, conocida como marcos de subcategorización. Por ejemplo, el verbo *dar* subcategoriza un grupo nominal y un grupo preposicional, en ese orden: *da un libro a María.*

El otro enfoque para describir formalmente la gramaticalidad de las oraciones es el de dependencias. Las dependencias se establecen entre pares de palabras, donde una es principal o rectora y la otra está subordinada a (o dependiente de) la primera. La única palabra que no está subordinada a otra es la raíz del árbol.

La motivación de muchas dependencias sintácticas es el sentido de las palabras. Por ejemplo en la frase *Los niños pequeños estudian pocas horas*, las palabras *pequeños* y *pocas* son modificadores de atributo de las palabras *niños* y *horas* respectivamente, y *niños* es el sujeto de *estudiar*. Un rasgo muy importante de las dependencias es que no son iguales: una sirve para modificar el significado de la otra, así la secuencia *los niños pequeños* denota ciertos niños, y *estudian pocas horas* denota una clase de estudio.

En el enfoque de dependencias, la línea de trabajo más importante es la desarrollada por el investigador Igor Mel'cuk, la *Meaning ⇔ Text Theory* (MTT). Para (Mel'cuk, 79), en la sintaxis se describen los medios lingüísticos por los cuales se expresan todos los participantes que están implicados en el sentido mismo de los lexemas[1], es decir, sus valencias. Por ejemplo, el lexema *plática* indicará que utiliza la preposición *sobre* para introducir el tema, que *solidaridad* utiliza la preposición *con*, y que el verbo *dar* emplea un sustantivo para expresar el objeto donado y para introducir el receptor emplea la preposición *a*.

Para caracterizar las valencias sintácticas del español nos basamos en la comparación de los enfoques existentes, y principalmente en la MTT, donde con la ayuda de una tabla de patrones de manejo[2] (*Government Patterns*, GP) (Steele, 90), se relacionan los participantes semánticos o actantes con los complementos del lexema encabezado, es decir, la información de correspondencia entre valencias semánticas y sintácticas. Los GP describen también todas las formas en que se realizan las valencias sintácticas y la indicación de obligatoriedad de la presencia de cada actante, si es necesario. Después de la tabla de GP se presentan dos secciones: restricciones y ejemplos. Las restricciones consideradas en los GP son de todo tipo: semánticas, sintácticas o morfológicas. La sección de ejemplos cubre todas las posibilidades: ejemplos para cada actante, ejemplos de todas las posibles combinaciones de actantes y finalmente los ejemplos de combinaciones imposibles o indeseables.

La parte principal de la tabla de GP es la lista de valencias sintácticas del lexema encabezado. Se listan de una manera arbitraria pero se prefiere el orden de incremento en la oblicuidad: sujeto, objeto directo, objeto indirecto, etc. También la forma de expresión del significado[3] del lexema encabezado influye en el orden, por ejemplo la expresión para *acusar*: Person V accuses person W in action X. Esta expresión precede cada GP.

Otra información obligatoria en cada valencia sintáctica es la lista de todas las posibles formas de expresión de la valencia en los textos. El orden de opciones para una valencia dada es arbitraria, pero las opciones más frecuentes aparecen normalmente primero. Las opciones se expresan con símbolos de categorías gramaticales y palabras específicas.

A continuación presentamos una descripción para el

1 Unidad básica del léxico, portadora de significado propio

2 Una traducción más adecuada para este término sería *Patrones de Rección*, para evitar la confusión con la misma palabra empleada en la Teoría de la Rección y el Ligamento de N. Chomsky, elegimos *manejo sintáctico*.

3 Empleamos el inglés para la descripción de significado puesto que no existe un lenguaje semántico sin homonimia ni sinonimia, por lo que el inglés parece más conveniente que el mismo español para hispanohablantes.

verbo *acusar,* una descripción más amplia de este diccionario aparece en (Galicia et al, 98). En esta descripción no se presenta en forma exhaustiva la sección de ejemplos; en la descripción NP representa un grupo nominal e INF representa un verbo en infinitivo.

1 = V	2 = W	3 = X
1. NP	2. *a* NP	1. *de* NP
		2. *de* INF
Obligatoria	Obligatoria	

Posibles

C.1 + C.2	La policía *acusa* a Ana.
C.1 + C.2 + C.3.1	La policía *acusa* a Ana de robar.

Prohibidas:

C.1 + C.3.1	La policía *acusa* de robar.
C.3.1	*Acusa* de robo.

2.2 Características del Español

Orden de palabras. El orden de palabras en el español es más libre comparado con el inglés y por lo tanto se requiere considerar los órdenes posibles de aparición de las valencias. Por ejemplo, en las frases siguientes el objeto indirecto no aparece después del verbo, de tres maneras distintas: 1) en la forma *a* NP antes del verbo, 2) como pronombre reflexivo entre sujeto y verbo, y 3) como clítico dentro del verbo.

1. *A quienes acusan de comportamiento arrogante.*
2. *El fiscal me acusa de delito de alta traición.*
3. *Acusándole de ser el sostenedor y portavoz de Mario Segni.*

Para el inglés funciona buscar usualmente todos los objetos del verbo después de él. Sin embargo, para el español, esta información de posibles posiciones de la valencia es necesaria para el analizador sintáctico. En la siguiente sección se presentan todas las combinaciones obtenidas del corpus LEXESP[4] para el verbo *acusar.*

Sujeto y objeto directo. La inversión del sujeto es considerada como un recurso estilístico de frecuencia de aparición menor, pero a este respecto, el español presenta diferencias con otros lenguajes romances: el español y el italiano permiten la inversión libre del sujeto, a diferencia del francés. Por ejemplo, en las siguientes frases, el sujeto aparece después del verbo en dos formas distintas, como nombre propio y como grupo nominal.

1. *Le acusaba Apel de desembocar en una ilusión idealista por <...>.*
2. *A quien acusaron varios testigos.*

Considerando un orden de palabras fijo, podría haber un reconocimiento erróneo de valencias o del significado del verbo o ambos. En la primera frase sólo si no se reconoce el nombre propio. En la segunda frase se requiere diferenciar entre entidad animada y grupo nominal inanimado para reconocer el sujeto de $acusar_1$ 'denunciar a alguien como culpable de algo'. La valencia realizada como NP inanimado corresponde al verbo $acusar_2$ 'revelar algo, ponerlo de manifiesto'. Si *varios testigos* se reconoce como NP inanimado existe confusión entre sujeto y complemento, que resultará en una asignación de estructura incorrecta o de otro significado.

En la mayoría de los lenguajes el objeto directo está conectado con el verbo sin preposiciones; pero en español, las entidades animadas están conectadas con la preposición *a* y las no animadas directamente (*veo a mi vecina* y *veo una casa*). La animidad se considera como una personificación, por ejemplo *gobierno* en español es un sustantivo animado y al dirigirse a él se utiliza la preposición *a* (*veo al gobierno...*). Además de personas, la animidad abarca grupos de personas, animales, países, entidades abstractas (organizaciones, partidos políticos), etc. En cambio en ruso donde también existe oposición obligatoria de palabras por animidad, los grupos de personas, los países, las ciudades no se personifican en sentido gramatical.

Aunque la preposición *a* también tiene otros usos, nos referimos exclusivamente a su conexión con el objeto directo. Este uso sirve para diferenciar el significado de algunos verbos, por ejemplo, *querer algo* 'tener el deseo de obtener algo' y *querer a alguien* 'amar o estimar a alguien'. Así que la animidad es una característica evidentemente sintáctica pero con alusión semántica que se considera para la realización de las valencias y en ciertos casos permite determinar el sujeto y distinguirlo del objeto.

Valencias repetidas. Generalmente las entidades referidas por diversas valencias son diferentes. Esta es una situación normal en lenguajes naturales, cada valencia semántica puede representarse en el nivel sintáctico mediante un solo actante. Sin embargo, existen lenguajes como el español que permite la duplicación de valencias. Los siguientes ejemplos muestran en cada oración dos grupos en negritas que representan el mismo objeto:

- *Arturo* ***le*** *dio la manzana* ***a Víctor.***
- ***El disfraz de Arturo, lo*** *diseñó Víctor.*
- ***A Víctor le*** *acusa el director.*

Mientras que en la primera oración se duplica el objeto indirecto, en las siguientes oraciones se duplica el objeto directo. Mientras en la primera oración la repetición es opcional, en las siguientes es obligatoria. El orden de palabras y los verbos específicos imponen algunas construcciones; por ejemplo, en las oraciones anteriores, el cambio de los argumentos dativos y acusativos antes del verbo.

[4] El corpus LEXESP nos fue proporcionado amablemente por H. Rodríguez de la Universidad Politécnica de Cataluña, en Barcelona, España.

2.3 Patrones de Manejo Avanzados

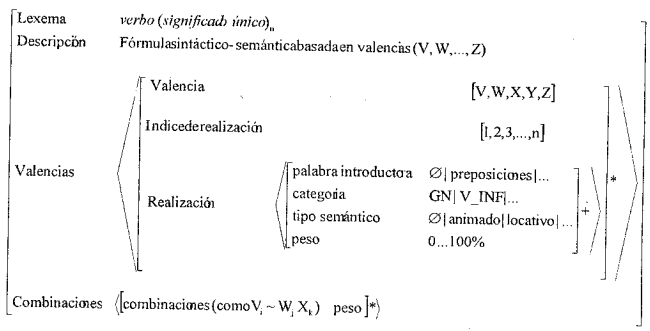
Proponemos una nueva estructura de GP que une los méritos de los enfoques existentes, la llamamos Patrones de manejo sintáctico avanzados (PMA), ver Figura: 2, que además de un formato modernizado para sistemas computacionales, incluye nuevos atributos para algunas características del español y probabilidades para la realización y compatibilidad de valencias. La información contenida en estos patrones corresponde a la expuesta anteriormente, considerada en la tabla de GP, salvo la indicación de obligatoriedad de la presencia de cada valencia. En un PMA, la indicación de obligatoriedad, las posibles combinaciones de actantes y las combinaciones prohibidas han sido consideradas de otra forma.

El español tiene un orden de palabras más libre que el inglés pero no totalmente libre, por lo que las posibles combinaciones de valencias son limitadas. A partir de la indicación de obligatoriedad se pueden definir algunas combinaciones no deseadas pero no la totalidad. Las combinaciones posibles y las prohibidas pueden definirse basándose en cierta experiencia pero no reflejarían los cambios en el lenguaje ni las preferencias en dominios específicos. Por lo que para especificar esta información consideramos la obtención de pesos estadísticos. Si una valencia tiene presencia en todas las oraciones extraídas del corpus para un verbo específico, se considera como una evidencia de obligatoriedad. El analizador sintáctico empleará esta evidencia para buscar las valencias aún en posiciones distantes. Por ejemplo, el verbo *acusar* requiere la presencia del objeto directo, con esta indicación el analizador sintáctico buscará este pedazo de información alrededor del verbo, considerando también las probabilidades de su aparición antes y después del verbo.

Así que obtenemos los pesos estadísticos para cada valencia referidos a las palabras introductoras de ellas y después los pesos estadísticos de las combinaciones de valencias referidas a la posición de cada valencia respecto al verbo. Esta información estadística da un rango de las descripciones de cada tipo específico de cada valencia y de sus combinaciones, que permitirá incrementar la eficiencia del analizador sintáctico.

En las Figuras 2 y 3 se muestran, en una presentación más práctica que la definida en la Figura 1, los PMA obtenidos a partir de un total de 227 oraciones del corpus LEXESP para el verbo *acusar*. De la información obtenida, se reconocen $acusar_1$ y $acusar_2$. La valencia realizada mediante NP en todos los casos marca la diferencia entre los dos verbos, siempre y cuando pueda discriminarse entre entidades animadas (*an*) y no animadas. Nótese en estas figuras las diversas variantes en el orden de palabras, y la representación de valencias duplicadas en la Figura 1 con el caso $[W_2V \sim XW_2, 1.76\%]$.

3 Análisis Sintáctico y Desambiguación



Donde: + denota uno o más elementos * denota cero o más elementos
~ denota el verbo

Figura: 2 Patrones de manejo sintáctico avanzados.

En el análisis sintáctico es necesario tratar con diversas formas de ambigüedad. La ambigüedad principal ocurre cuando la información sintáctica no es suficiente para hacer una decisión de asignación de estructura. La ambigüedad existe aún para los hablantes nativos, es decir, hay diferentes lecturas para una misma frase. Por ejemplo, en la oración *Javier habló con el profesor del CIC*, puede pensarse en *el profesor del CIC* como un complemento de *hablar* o también puede leerse que *Javier habló con el profesor* sobre un tema, habló con él *del CIC*.

Relacionada a la sintaxis, existe ambigüedad en el marcaje de partes del habla, esta ambigüedad se refiere a que una palabra puede tener varias categorías sintácticas, por ejemplo *ante* puede ser una preposición o un sustantivo, etc. Conocer la marca correcta para cada palabra de una oración ayudaría en la desambiguación sintáctica, es decir, en la selección de estructuras correctas, sin embargo la desambiguación de este marcaje requiere a su vez cierta clase de análisis sintáctico.
También existe ambigüedad en los complementos circunstanciales. Por ejemplo, en la frase *Me gusta beber licores con mis amigos*, el grupo *con mis amigos* es un complemento de *beber* y *no de licores*. Mientras un hablante nativo no considerará la posibilidad del complemento *licores con mis amigos,* para la computadora ambas posibilidades son reales.

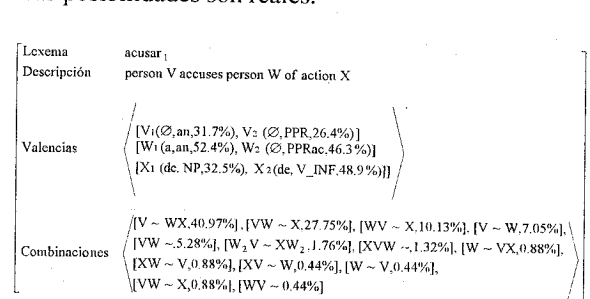


Figura 1 PMA del verbo $acusar_1$

Debido a que existe ambigüedad aún para los humanos, no es una tarea de la resolución de ambigüedades lograr una única asignación de estructuras en el análisis sintáctico de textos, sino eliminar la gran cantidad de variantes erróneas que normalmente se producen. El esquema de análisis

Lexema	$acusar_2$
Descripción	V reveals W
Valencias	⟨[V,⟨(∅, NP,24.3%),(∅, an,37.8%),(∅, PPR,16.2%)⟩] ⟨[W,⟨(∅, NP,100%)⟩]
Combinaciones	⟨[V ~ W, 97.3%],[WV ~, 2.7%]⟩

Figura: 3 PMA del verbo $acusar_2$

sintáctico y desambiguación que proponemos considera tres fuentes de conocimiento: léxica, sintáctica y semántica (ver Figura: 4). Sólo con la participación de estos conocimientos es posible diferenciar las variantes sintácticas correctas de entre las múltiples variantes sintácticas obtenidas en el proceso de análisis sintáctico. Cada módulo proporciona un conjunto de variantes con pesos basados en características satisfechas de cada método. Así que la salida de cada módulo da una medida cuantitativa de la probabilidad de cada estructura sintáctica. Mediante esta medida, un módulo de votación clasifica las variantes para que en el tope aparezcan las más probables de ser las correctas.

El modelo general se presenta en la Figura: 4. Dado el carácter cuantitativo del modelo, a futuro pueden incluirse otros métodos. Las fuentes de conocimiento consideradas actualmente son:

1. *Patrones de manejo,* que reflejan conocimiento léxico y sintáctico.

Este método se basa en conocimiento lingüístico que adquieren los hablantes nativos durante el aprendizaje de su lengua, por lo que se considera el método principal. El conocimiento descrito en este módulo es la información léxica de verbos, adjetivos y algunos sustantivos del español, para enlazar las frases que realizan las valencias, es un diccionario de PMA. La compilación de este tipo de diccionarios, hasta ahora solamente ha sido posible manualmente, por lo que su cobertura ha sido limitada. No es posible establecer ese conocimiento mediante reglas o algoritmos pero es posible obtener la información léxica a partir de un corpus, para lo cual proponemos un método descrito en la siguiente sección.

El peso asignado a cada variante depende del número total de patrones y de valencias empatados, así como del tipo de patrones considerados, de las frecuencias de realización de las valencias y del número de homónimos en los patrones.

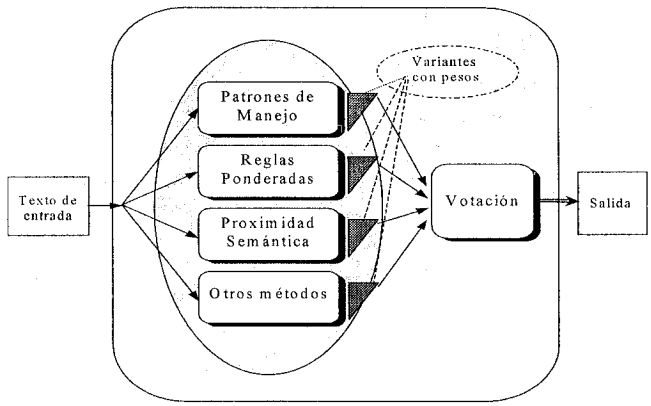


Figura: 4 Modelo de desambiguación.

2. *Reglas ponderadas*, que reflejan conocimiento sintáctico.

El método de reglas ponderadas se basa en una gramática de constituyentes extendida con rasgos de concordancia (género, número, persona). La gramática de constituyentes que creamos se apoya en las marcas morfológicas que contienen las palabras de LEXESP. Para hacer del método la herramienta básica de análisis sintáctico introdujimos varios elementos. La inclusión del elemento rector en cada regla nos permite hacer una transformación de estructura de constituyentes a estructura de dependencias. La inclusión de relaciones sintácticas, además de establecer la dirección de las dependencias, permite diferenciar entre valencias y algunos complementos circunstanciales.

Otros elementos introducidos, como los elementos de puntuación, las marcas semánticas de tiempo, y principalmente los pesos en las reglas, incrementan la calidad del análisis mismo. Los pesos introducidos en las reglas permiten graduar el número de reglas que se usan en el análisis, de esta forma se da mayor prioridad a las construcciones más usuales. Incluimos dos niveles de detalle, un nivel general donde se van aplicando prioridades (primero aplican las reglas de mayor prioridad y si no es posible el análisis total se continúa con el siguiente grupo de reglas con menor prioridad) y un nivel de detalle en nodos interiores para utilizar las reglas que tengan mayor prioridad para las mismas subestructuras. La gramática que necesitamos en este caso, dado que no es el método más importante, no requiere condiciones óptimas en cuanto a cobertura y precisión. Nuestra gramática pretende considerar las construcciones más comunes.

La labor requerida para realizar la clasificación y la asignación de valores, comparada contra los resultados de un método que no distingue información léxica y da estructuras iguales por categorías gramaticales nos hizo proponer una asignación de pesos por igual para todas las variantes, con la finalidad de que los métodos de PMA y de proximidad semántica sean los que hagan emerger las variantes correctas.

3. *Redes semánticas*, que reflejan conocimiento semántico de cercanía de sentido entre grupos sintácticos.

La información léxica puede ayudar a resolver muchas ambigüedades pero en otros casos es necesaria la proximidad semántica para la desambiguación. Por ejemplo: *Me gusta beber licores con menta* y *Me gusta beber licores con mis amigos;* en ambas frases la clase semántica del sustantivo final ayuda a resolver la ambigüedad, es decir con que parte de la frase están enlazadas las frases preposicionales, *con menta* y *con mis amigos*. Ni *menta* ni *amigos* son palabras ambiguas pero *amigos* está más cercana semánticamente a *beber* que a *licores* y *menta* está más cercana a *licor* que a *beber*.

El empleo de la red semántica para la desambiguación sintáctica tiene la finalidad de incorporar la componente semántica, conocimiento de contexto local, faltante en las otras dos fuentes. Cuando varias estructuras son igualmente posibles o el enlace de adjuntos (complementos circunstanciales no relacionados al significado de la palabra a la que se enlazan) es ambiguo, la proximidad semántica puede ayudar, es decir, los conceptos más cercanos relacionados a las palabras en los constituyentes posibles.

Algunas de las gramáticas más actuales, derivadas de las gramáticas generativas precisamente incorporan restricciones semánticas, como la HPSG (Sag & Wasow, 99), que las considera en la entrada de cada lexema en el diccionario, lo cual implica una labor manual, intensiva en extremo. En nuestro modelo, esas restricciones semánticas se buscan en la red y se definen a través de la proximidad semántica. Aunque las redes semánticas son una aproximación a las habilidades humanas y por lo tanto son modelos simplificados, pueden usarse de una forma acorde a sus limitaciones.

Crear una red semántica es una tarea de labor intensa, y difícil de lograr aún a largo plazo. En este trabajo consideramos la red semántica que se está desarrollando a partir de la red FACTOTUM[5], mediante un método de traducción (Gelbukh, 98). Para resolver la ambigüedad sintáctica, los enlaces de palabras o de grupos de palabras se realizan determinando el grado de proximidad semántica que tienen esas palabras o grupos de palabras.

La determinación de la proximidad semántica se basa en las características de la red semántica, que son: conceptos, relaciones, y trayectorias. Al igual que otros autores, describimos la proximidad semántica como un valor cuantitativo pero para determinarla no solamente consideramos la longitud por el número de enlaces sino también un peso asignado de acuerdo al tipo de relación. La trayectoria misma representa un valor cualitativo.

La proximidad entre un par de palabras es un valor que depende de la longitud y del tipo de relación. Para nosotros depende: 1) del tipo de relación, 2) de los enlaces individuales y 3) de las relaciones implícitas. La primera

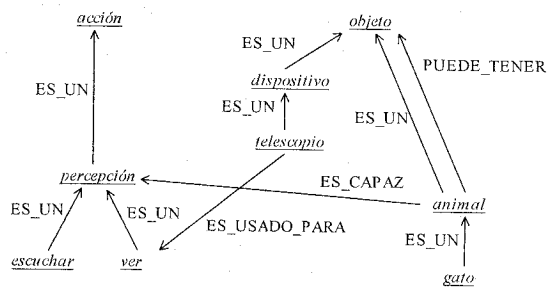


Figura 6: Fragmento de la red semántica para la frase *Veo un gato con un telescopio.*

asignación contempla los valores mismos de las relaciones explícitas, es decir, su importancia. La segunda asignación pretende corregir el problema que se presenta conforme las relaciones están más cercanas al tope de la jerarquía, mientras más alejadas del tope, las palabras tienen más aspectos comunes. La tercera asignación considera la problemática de las inferencias. Por ejemplo la relación *carro* ES_UN *objeto* y la relación implícita *objeto* TIENE_SUBTIPO *libros*. De esta forma, la trayectoria es corta a pesar de que no hay muchos aspectos comunes. Para resolver este problema se asigna un peso mayor a una relación implícita que a una explícita. La precisión se obtiene junto con la segunda asignación que hace mayor la longitud de *carro* ES_UN *objeto* que de *Ford* ES_UN *carro*.

Desambiguación sintáctica. En el empleo de la red semántica para la desambiguación sintáctica realmente se está incorporando la componente semántica faltante en las reglas ponderadas. Por lo que la evaluación de la proximidad no sólo está relacionada con los valores obtenidos de la red misma, sino que es necesario considerar además el tipo sintáctico de la relación. No todas las trayectorias son aceptables en un contexto específico. En algunos casos se tendrá que buscar la trayectoria con las relaciones que sean más adecuadas al contexto sintáctico de la oración. Por ejemplo, en la frase *Veo un gato con un telescopio* aparece la frase preposicional *con un telescopio*, la relación más cercana será USO y una relación más cercana, tipo ES_UN, no será la más adecuada para ese contexto (Figura 6).

Así que la tarea de desambiguación está muy relacionada con el método para encontrar las trayectorias aceptables mínimas y de contexto sintáctico. En una red semántica existe un número infinito de trayectorias conectando dos palabras

Módulo de votación. Para desambiguar las estructuras sintácticas un módulo de votación emplea los pesos asignados en cada módulo, vota por el máximo valor sumado de las variantes. El resultado es una lista clasificada de las variantes sintácticas.

Para poder hacer la votación entonces, nuestro modelo

[5] *FACTOTUM® SemN et*, es una red semántica compilada por la empresa MICRA, INC. New Jersey, USA.

requiere una evaluación cuantitativa para ordenar las

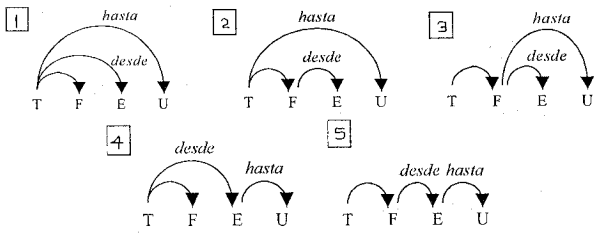


Figura 7: Variantes de la estructura sintáctica[6] para la frase *Trasladaron la filmación desde los estudios hasta el estadio universitario.*

variantes construidas por cada módulo, y una forma que las haga compatibles para su evaluación. La compatibilidad se logra mediante un formato común que es la estructura de dependencias. Se transforman las estructuras de constituyentes a una estructura de dependencias, para hacer posible la comparación de sus valores con los valores de las estructuras del módulo de patrones de manejo.

A continuación presentamos un ejemplo, para la frase *El productor trasladó la filmación de los estudios al estadio universitario*, indicando solamente las ideas del modelo. En este ejemplo consideramos lo siguiente:

- Patrones de manejo:

4.34896 *trasladar*,dobj_suj,obj:a,obj:de, x:?
0.436967 *trasladar*, obj:a, x:?
1.13758 *filmación*
3.29976 *estadio*

donde los números de la primera columna representan los valores obtenidos del método de compilación de información sintáctica para los patrones de manejo. La marca "x" representa una valencia repetida mediante clíticos, la marca "?" representa NP, "dobj_suj" indica el grupo nominal que puede ser sujeto u objeto directo, "obj" indica los complementos preposicionales y enseguida la preposición específica.

Así que considerando los patrones de *trasladar, estadio* y *filmación* se favorecen las variantes con la estructura de: 'trasladar algo a algún lugar desde otro lugar'.

- Con el modelo de reglas ponderadas obtenemos 8 variantes con el mismo peso.
- Con la proximidad semántica, encontramos las siguientes relaciones:

filmación -> *director*
-> subtipo de *espectáculo*
trasladar -> con referencia a una dirección o a un lugar
-> con relación a traslación de un objeto
-> subtipo trayectoria
estudio cinematográfico > lugar
estadio -> como subtipo de *espectáculo*
filmación -> *cine* -> *director*

Únicamente la relación entre *trasladar* y *estudio* como *lugar* puede considerarse, favoreciendo la estructura: 'trasladar desde lugar' (*trasladar de los estudios*). En este ejemplo el método de patrones de manejo es el que más contribuye para reconocer las variantes correctas.

Para valorar la precisión de los métodos desarrollados para desambiguación se han considerado diversos elementos de evaluación. Por ejemplo, enlaces entre palabras de contenido (Yuret, 98) comparados contra un conjunto de oraciones analizadas manualmente, o el número de constituyentes obtenidos (Collins, 99) comparado contra un corpus marcado sintácticamente. En nuestro caso consideramos la aparición de las variantes más probables de ser las correctas en un rango en el tope de la clasificación, utilizando un conjunto de oraciones analizadas manualmente.

4 Colección de Estadísticas de las Combinaciones de Subcategorización

La información requerida para llenar los PMA es la información usual de los marcos de subcategorización más la relación de esta información sintáctica con las valencias, las características semánticas (animado, locativo), y las estadísticas de uso común.

Para el español, no hay diccionarios con información completa de subcategorización y sólo existen algunos intentos para la adquisición automática de marcos de subcategorización. Además, tomando en cuenta que el español tiene restricciones menos estrictas en el orden de palabras, la combinación de complementos resulta en un problema mayor. Por lo que proponemos un método de desambiguación para obtener la información. El método se basa en las estadísticas de los errores que un analizador sintáctico específico empleado para el análisis de textos hace en unos documentos determinados. Para cada frase, se determina un *peso* (o probabilidad) para cada variante.

Como ejemplo presentamos la frase: *Trasladaron la filmación desde los estudios hasta el estadio universitario.* A esta frase puede asignársele al menos las interpretaciones sintácticas mostradas en la Figura 7 donde aparecen árboles de dependencias simplificados para cinco variantes. Para este ejemplo, un hablante nativo escogería la primera estructura como la interpretación más probable, tomando en cuenta cierta información léxica. El analizador sintáctico requiere esa información léxica para desambiguar la frase.

La selección de estructura se realiza en términos cuantitativos, asignando un peso o una probabilidad a cada variante; a mayor peso mayor la probabilidad de que la

[6] Las letras significan: T = trasladar, F = la filmación, E = estudios y U = estadio universitario.

variante sea la correcta. Trabajamos con las estadísticas de marcos de subcategorización individuales a los que llamamos *combinaciones*. Por ejemplo el árbol número 1 de la Figura 7 contiene solamente una combinación: *trasladar +desde +hasta +?*, el árbol número 3 contiene: *trasladar +?, filmación +desde +hasta*. La selección de estas combinaciones no es aleatoria, en buen grado realmente son fijas para cada palabra, así que sus estadísticas son más confiables que las de pares de palabras arbitrarias .

La desambiguación propuesta se basa en las frecuencias de esas combinaciones, en las variantes correctas (p_i^+) y en los árboles generados por el analizador pero rechazados por revisores (p_i^-). No es un gran problema calcular esos pesos si se cuenta con texto marcado sintácticamente. Sin embargo, para un área específica, para un género determinado o para un lenguaje como el español o el ruso, no se dispone de esa información.

En nuestro método hay dos metas interrelacionadas, primero determinar la estructura correcta de cada frase y segundo, compilar un diccionario para la desambiguación. El procedimiento que usamos es iterativo. Aproxima las dos metas mencionadas en pasos alternados: primero estima las hipótesis en base a los pesos actuales en el diccionario, después reevalúa los pesos en el diccionario conforme a los pesos de las hipótesis de cada frase[7], y repite el proceso.

El proceso comienza con un diccionario vacío. En la primera iteración, para cada frase, todas las hipótesis producidas por el analizador sintáctico tienen pesos iguales. Después, se determinan las frecuencias p_i^+ y p_i^- para cada combinación encontrada al menos una vez en cualesquiera de las variantes producidas por el analizador. Ya que se desconocen las variantes correctas, para calcular p_i^+ se suman los pesos w_j de cada variante donde la combinación fue encontrada, porque estos pesos representan la probabilidad de que la variante sea correcta. Similarmente, para determinar p_i^- sumamos los valores (1 - w_j) que representan la probabilidad de que la variante dada sea incorrecta. Lo anterior se resume en estas expresiones:

$$p_i^+ = \sum w_j / S,$$
$$p_i^- = \left[\sum \left(1 - w_j\right) + \lambda\right] / (V - S),$$
$$w_j = C \times \prod \left(p_i^+ / p_i^-\right), \quad \sum w_k = 1,$$

donde S es el número total de oraciones, V es el número total de variantes (hipótesis) en el corpus, C es una constante de normalización y λ es un valor para contrarrestar los efectos de palabras que no existían previamente en los datos. Una explicación detallada de la obtención de estas fórmulas se presenta en (Gelbukh *et al*, 98).

Para nuestros propósitos, creamos una gramática extendida independiente del contexto (con concordancia) e implementamos un analizador sintáctico tipo *chart*. La Tabla 1 muestra los resultados producidos por el algoritmo,

+/-	Combinación
11.3512	obj:con,obj:de,x:?
11.3512	dobj_suj:?,obj:de,x:?
11.3512	dobj_suj:?,obj:con,obj:de,x:?
11.3512	dobj_suj:?,dobj_suj:?,obj:de,x:?
11.3512	acusar,dobj_suj:?,dobj_suj:?,obj:con, obj:de,x:?
4.48254	obj:de,x:?
3.59065	obj:de,obj:de,x:?
3.00859	x:?
1.32416	dobj_suj:?,dobj_suj:?,obj:a,obj:de
1.29413	dobj_suj:?,dobj_suj:?,obj:a
0.691213	dobj_suj:?,obj:a,obj:de
0.620666	dobj_suj:?,obj:a
0.408024	obj:a,obj:de
0.378879	obj:a,x:?
0.378879	obj:?,obj:a,x:?

Tabla 1: Resultados del método estadístico.

para las combinaciones del verbo *acusar,* cuando se aplica al corpus LEXESP. A pesar de que estos resultados fueron obtenidos con muy pocas oraciones, podemos comparar con la sección 2.1 y solamente 3 combinaciones son incorrectas, las que tienen "obj:con" que consideran un objeto introducido con la preposición *con*. Estos errores se eliminarían al analizar mayor número de oraciones. En la tabla 1, el valor p_i^+/p_i^- es representado por +/-, dobj_suj indica objeto directo o sujeto. No hay un orden, por lo que *acusar, dobj_suj:?, obj:?* también representa *acusar, obj:?,dobj_suj:?*

El algoritmo es no supervisado y produce una lista de preposiciones usadas con cada palabra. La técnica para adquirir este conocimiento léxico es diferente de los métodos conocidos para enlaces de grupos preposicionales, porque se dirigen al enlace de patrones tipo, como por ejemplo V N1 P N2 en (Ratnaparkhi, 98), o emplean textos con marcas sintácticas como en (Merlo et al, 97).

5 Conclusiones

La principal contribución de este trabajo es en el avance del análisis sintáctico de textos en español sin restricción. En el español, la ambigüedad sintáctica se ve magnificada por la cantidad de frases preposicionales que se emplean, lo que ocasiona una mayor cantidad de variantes obtenidas en el análisis sintáctico. Este problema es el que se disminuye con nuestro modelo.

En experimentos realizados, los resultados obtenidos en la aplicación de nuestro método de compilación de combinaciones de subcategorización al corpus LEXESP fueron usados para analizar sintácticamente 100 oraciones y las estructuras verdaderas se encontraron clasificadas en el rango tope del 35%.

[7] El peso de la variante es el producto de los pesos de sus combinaciones.

Referencias

Collins, M. *Head-driven Statistical Models for Natural language parsing.* Ph.D. Thesis University of Pennsylvania. 1999. http:// xxx.lanl.gov/ find/cmp-lg/

Galicia Haro, S., I. Bolshakov, A. Gelbukh. *Diccionario de patrones de manejo sintáctico para análisis de textos en español.* Revista SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural, No. 23, España, pp.171-176. Septiembre de 1998.

Gelbukh, A. F. *Lexical, syntactic, and referencial disambiguation using a semantic network dictionary.* Technical report. CIC, IPN, 1998.

Mel'cuk, I. A. *Dependency Syntax.* In P. T. Roberge (ed.) Studies in Dependency Syntax. Ann Arbor: Karoma 23-90, 1979.

Merlo, P., Crocker, M. and Berthouzoz, C. *Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation.* In Proceedings of the EMNLP-2, 1997. http://xxx.lanl.gov/find/cmp-lg/9710005

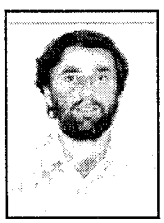
Ratnaparkhi, A. *Statistical Models for Unsupervised Prepositional Phrase Attachment.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Quebec, Canada, 1998 http://xxx.lanl.gov/ps/cmp-lg/9807011

Sag, I. A. and Wasow, T. *Syntactic Theory: A Formal Introduction.* CSLI 1999.

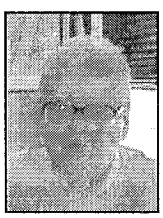
Steele, J. *Meaning - Text Theory. Linguistics, Lexicography, and Implications.* James Steele, editor. University of Ottawa press, 1990.

Yuret, D. *Discovery of Linguistic Relations Using Lexical Attraction.* Ph. D. thesis. Massachusetts Institute of Technology. 1998. http:// xxx.lanl.gov/ find/cmp-lg/9805009

Sofía N. Galicia Haro, *es Ingeniero en Comunicaciones y Electrónica de la ESIME-IPN. Obtuvo el grado de Maestra en Ciencias de la Computación en el Instituto de Investigación en Matemáticas Aplicadas y Sistemas de la Universidad Nacional Autónoma de México en 1994, en donde ha sido profesor de asignatura. Laboró en Luz y Fuerza, empresa que la apoyó para realizar estudios de doctorado. En el 2000 obtuvo el grado de Doctor en Ciencias de la Computación en el Centro de Investigación en Computación del Instituto Politécnico Nacional. Es autora de alrededor de 20 publicaciones en Lingüística Computacional. Actualmente realiza actividades posdoctorales en el Laboratorio de Lenguaje Natural del CIC-IPN y es profesor de asignatura en la Facultad de Ciencias de la UNAM. Desde Julio del 2002 es Investigador Nacional nivel 1 del Sistema Nacional de Investigadores de CONACyT.*



Alexander Gelbukh, *nació en Moscú, Rusia, en 1962. Obtuvo el grado de Maestro en Ciencias en Matemáticas en 1990 de la facultad de Mecánica y Matemáticas de la Universidad Estatal «Lomonósov» de Moscú, Rusia, y el grado de Doctor en las Ciencias de la Computación en 1995 del Instituto de la Información Científica y Técnica de Toda Rusia (VINITI), Rusia. Desde 1997 es el jefe del Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación del Instituto Politécnico Nacional, México. Es miembro de la Academia Mexicana de Ciencias desde 2000, del Sistema Nacional de Investigadores desde 1998, conferenciante de Programa de Conferencias para México y Centroamérica de la ACM desde 2000; autor, coautor o editor de más de 200 publicaciones en la lingüística computacional; fundador y organizador de la serie CICLing de congresos en la lingüística computacional, véase www.CICLing.org. Para mayor información véase www.Gelbukh.com ó www.cic.ipn.mx/~gelbukh.*



Igor A. Bolshakov, *nació en Moscú en 1934. Obtuvo el grado de Maestría en Física en 1956 por el Departamento de Física de la Universidad estatal de Moscú "Lomonossov", Rusia. Obtuvo el grado de Doctor en 1961 por el Instituto VYMPEL de Moscú, Rusia y el grado de Doctorado científico en Ciencias de la computación en 1966 por el mismo Instituto. Recibió el Premio Nacional de la USSR en Ciencia y Tecnología en 1989. Desde 1996 trabaja en el Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación del Instituto Politécnico Nacional. Es miembro del SNI, México con nivel II desde el año 2000. Es autor de alrededor de 200 publicaciones en teoría de radares, teoría de probabilidades, y en lingüística computacional.*