Extracción automática de un diccionario de colocaciones en español

Sulema Torres-Ramos

Centro de Investigación en Computación (CIC-IPN),
Unidad Profesional Adolfo López Mateos,
Av. Juan de Dios Bátiz s/n esquina M. Othón de Mendizábal,
Zacatenco, México, D.F. 07738, México.
sulema7@gmail.com

Resumen. Las colocaciones son pares de palabras de contenido que forman las relaciones sintácticas de dependencia razonables, directamente o a través de palabras funcionales. Tales pares tienden usarse en los textos más frecuentemente de lo esperado por casualidad. El texto en lenguaje natural consiste casi totalmente de tales colocaciones. La información de las palabras que forman colocaciones es útil en diferentes aplicaciones de procesamiento de lenguaje natural. Actualmente existen corpus etiquetados con estructura sintáctica mismos que pueden ser útiles para la extracción de colocaciones.

En este artículo se presenta la extracción automática de un diccionario estadístico grande de colocaciones a partir de un corpus con las estructuras sintácticas marcadas manualmente. Las relaciones de dependencias encontradas en tal corpus, junto con sus frecuencias, constituyen nuestro diccionario de colocaciones.

Palabras clave: Lingüística Computacional, Procesamiento de Lenguaje Natural, Colocaciones, Ambigüedad Sintáctica.

1 Introducción

Una colocación es la relación entre dos palabras o un grupo de palabras que frecuentemente se usan de manera conjunta formando una expresión común [1]. Algunos ejemplos de colocaciones en español son *sistema político*, *seguro de vida*, *núcleo familiar*, etc.

Supongamos a un estudiante escribiendo un ensayo acerca del medio ambiente. Él conoce los temas que desea cubrir y tiene las ideas y los argumentos para hacerse entender. Además posee un repertorio de vocabulario útil, especialmente de sustantivos de alto contenido como *medio ambiente*, *contaminación*, *capa de ozono*. Lo que hace falta son las palabras que pueden ligar el vocabulario de alto contenido para convertirlo en un texto coherente -como un argumento o una narración-. La contaminación es un problema, pero ¿qué se necesita hacer al respecto? Buscando en un diccionario y examinando rápidamente en la sección de verbos nos arrojara las opciones de *evitar* / *prevenir* / *combatir* / *controlar* / *pelear* / *limitar* / *minimizar* / *reducir* /

monitorear. Con la ayuda de un diccionario común el estudiante puede escoger entre las opciones, la que exprese mejor lo que quiere decir. Sin embargo, si las colocaciones son difíciles de producir para un hablante no nativo [2,3], mucho más difícil lo es para una computadora.

Muchos esfuerzos se han hecho por generar manualmente diccionarios de colocaciones, sin embargo el tiempo y costo de esta tarea es muy alto. Actualmente existen corpus etiquetados con estructuras sintácticas que son la base para la extracción de colocaciones; de ahí surge la necesidad y relevancia de la obtención automática de colocaciones, utilizando los recursos existentes.

En este artículo se presenta un método automático para la extracción de colocaciones basado en un corpus etiquetado con las estructuras sintácticas. Específicamente, se llevó a cabo la extracción de un diccionario de colocaciones en español a partir del corpus etiquetado en español Cast3LB.

El artículo se organiza como sigue: primero, describimos a detalle las colocaciones, sus tipos y principales aplicaciones en lingüística computacional (sección 2) y se presentan los dos principales formalismos sintácticos en la sección 3. Después se presenta el método utilizado para la extracción del diccionario de colocaciones (sección 4). En la sección 5 se presenta la evaluación y resultados obtenidos. Al final se presentan las conclusiones y trabajo futuro.

2 Colocaciones

Hay mucha discusión y trabajo relacionado sobre colocaciones [4,5,6]. Dependiendo de los intereses y puntos de vista, los investigadores se enfocan en diferentes aspectos de las colocaciones.

Una de las definiciones más entendibles y usadas se encuentra en el trabajo lexicográfico presentado en [8]. La definición es la siguiente: una colocación es una combinación recurrente y arbitraria de palabras.

En [9] se define una colocación como: un par de palabras de contenido conectadas sintácticamente y que tienen compatibilidad semántica. Por ejemplo: *tomar* una *decisión*, *escuchar* la *radio*, *tocar* la *guitarra*, etc., donde los componentes de la colocación (colocativas) están subrayados.

Por palabras de contenido entendemos aquellas que tienen un significado, entre ellas se encuentran los sujetos, verbos, adjetivos, por ejemplo: perro, niño, comer, bonita, etc. Las palabras que no tienen contenido son los artículos como las, el, esos, etc.

La conexión sintáctica es entendida en las gramáticas de dependencias y ésta no es precisamente la coocurrencia de colocativas en un intervalo pequeño de texto. La colocativa rectora gobierna sintácticamente a la colocativa dependiente, estando adjunta a ella directamente o mediante una palabra auxiliar (usualmente una preposición). Secuencialmente, las colocativas pueden estar a cualquier distancia una de otra en una oración, mientras que en un árbol dependencias están muy cercanas.

2.1 Propiedades de las colocaciones

En esta sección, presentamos cuatro propiedades de las colocaciones que tienen relevancia en aplicaciones de lingüística computacional.

Las colocaciones son arbitrarias

Las colocaciones son difíciles de producir para un hablante no nativo [3]. No se trata simplemente de traducir palabra por palabra (word-for-word) lo que le gustaría al hablante decir en su lengua nativa. La tabla 1 muestra que la traducción palabra por palabra de "*to see the door*" corresponde en ambas direcciones de los cuatro lenguajes diferentes. Al contrario, traducir palabra por palabra la expresión "*to break down/force the door*" no tiene correspondencia en ambas direcciones en ninguno de los lenguajes.

La coocurrencia de "*door*" y "*see*" es una combinación libre, mientras que la combinación de "*door*" y "*break down*" es una colocación.

Para los hablantes no nativos de inglés es difícil construir correctamente la frase *"to break down a door"*.

Tabla 1. Comparaciones lingüísticas cruzadas de colocaciones

Lenguaje	Inglés	Traducción	Correspondencia en inglés
Francés	to see the door	voir la porte	To see the door
Alemán	to see the door	die Tür sehen	To see the door
Italiano	to see the door	vedere la porta	To see the door
Español	to see the door	ver la puerta	To see the door
Francés	to break down/force the door	enfoncer la porte	to push the door through
Alemán	to break down/force the door	die Tür aufbrechen	to break the door
Italiano	to break down/force the door	sfondare la porta	to hit/demolish the door
Español	to break down/force the door	tumbar la puerta	To fall the door

Traducir de un lenguaje a otro requiere más que buen conocimiento de estructura sintáctica y representación semántica, porque las colocaciones son arbitrarias, y deben ser fácilmente disponibles en ambos idiomas para que la traducción automática sea eficiente.

Las colocaciones son dependientes del dominio

Además de las colocaciones no técnicas tales como las que se presentaron antes, las colocaciones específicas del dominio son numerosas. Éstas son a menudo totalmente inentendibles para alguien ajeno al dominio. Contienen una gran cantidad de términos técnicos. Además, las palabras comunes se utilizan diferentemente. En el dominio de la navegación [10], por ejemplo, algunas palabras son desconocidas al lector no-familiar; la horca, y el sotavento son totalmente sin sentido para alguien ajeno a este dominio. Algunas otras combinaciones no contienen al parecer ninguna palabra técni-

ca, pero estas palabras adquieren un significado totalmente diferente en el dominio. Por ejemplo, un *traje seco* no es solamente un traje que está seco sino un tipo especial de traje usado por los marineros para permanecer seco en condiciones atmosféricas difíciles.

Dominar lingüísticamente un área específica requiere más que un glosario, requiere conocimiento de colocaciones dependientes del dominio.

Las colocaciones son recurrentes

La propiedad recurrente significa que las combinaciones de palabras no son excepciones, sino que se encuentran frecuentemente repetidas en un contexto dado.

Combinaciones de palabras como "tomar una decisión", "hacer un favor" son típicas del lenguaje, y colocaciones como "juntar hilos" son características de dominios específicos. Ambos tipos son frecuentemente usados en contextos específicos.

Las colocaciones son conjuntos de cohesión léxica

Por cohesión léxica [11] se entiende que la presencia de una o varias palabras de la colocación frecuentemente implica o sugiere el resto de la colocación. Esta propiedad es la más usada por lexicógrafos cuando compilan colocaciones [12,13].

Los lexicógrafos usan el juicio lingüístico de la gente para decir cuales son colocaciones y cuales no [14]. Ellos aplican cuestionarios a la gente, como el que se muestra en la siguiente tabla.

Tabla 2. Prueba *llenar-el-espacio* de Benson [8]

Oración	Candidatos
If a fire breaks out, the alarm will ??	ring/go off/ sound/start
The boy doesn't know how to ?? his bicycle	drive/ride/conduct
The American congress can ?? a presidential veto	ban/cancel/delete/reject
Before eating your bag of microwavable popcorn, you have to ?? it	cook/nuke/broil/fry/bake

Este cuestionario contiene las oraciones usadas por Benson para compilar el conocimiento de colocaciones para el diccionario BBI [15]. Cada oración tiene una ranura en blanco que puede ser fácilmente llenado por un hablante nativo (en este caso de inglés). En cambio, un hablante no nativo de inglés no encontraría las palabras faltantes automáticamente, sino que consideraría la lista de opciones de las palabras que tienen las características semánticas y sintácticas apropiadas, tales como las que están dadas en la segunda columna.

Como consecuencia, las colocaciones tienen una distribución estadística particular [5,16]. Esto significa que la probabilidad de que cualesquiera dos palabras adyacentes, por ejemplo, "arenque rojo" es considerablemente mayor que la suma de probabilidades de "rojo" y "arenque". Las palabras no pueden ser consideradas como variables independientes.

2.2 Tipos de colocaciones

Las colocaciones vienen en una gran variedad de formas. El número de palabras implicadas así como la forma de implicarlas puede variar mucho. Algunas colocaciones son muy rígidas, mientras otras son muy flexibles. Por ejemplo, una colocación compuesta por "tomar" y "decisión" puede aparecer como "tomar una decisión", "decisiones por tomar", "tomar una gran decisión", etc. En cambio, una colocación como "agente de ventas" puede aparecer sólo de una forma; esta es una colocación muy rígida, una expresión fija.

Se identifican tres tipos de colocaciones [11]: oraciones nominales rígidas, relaciones predicativas y plantillas de frase. A continuación se explican cada una de ellas.

Relaciones predicativas

Una relación predicativa consiste en dos palabras que se usan juntas repetidamente en una relación sintáctica similar [11]. Este tipo de colocación es la más flexible.

Por ejemplo, un sustantivo y un verbo formarán una relación predicativa si se usan juntos en varias ocasiones con el sustantivo como el objeto del verbo, "tomar-decisión" es un buen ejemplo de una relación predicativa. Así mismo, un adjetivo que frecuentemente modifica un sustantivo, como "niño-pequeño", es también una relación predicativa.

Esta clase de colocaciones se relaciona con las funciones léxicas de Mel'cuk [6,7], y las relaciones tipo L de Benson [15].

Oraciones nominales rígidas

Esta clase de colocaciones envuelve secuencias ininterrumpidas de palabras como "bolsa de valores", "procesamiento de lenguaje". Estas pueden incluir sustantivos y adjetivos, así como palabras de clase cerrada, y son similares al tipo de colocaciones recuperadas por [17,18]. Son el tipo más rígido de colocaciones. Algunos ejemplos son, "producto interno bruto", "impuesto al valor agregado", etc.

En general, las oraciones nominales rígidas no se pueden descomponer en fragmentos más pequeños sin perder su significado; son unidades léxicas en y de sí mismas. Por otra parte, frecuentemente se refieren a conceptos importantes en un dominio específico, y varias oraciones nominales rígidas se pueden utilizar para expresar el mismo concepto.

Plantillas de frase

Consisten en frases idiomáticas que contienen una, varias o ninguna ranura en blanco. Son colocaciones de frase largas. Algunas colocaciones de este tipo, en el dominio de la bolsa, se muestran a continuación:

*En la bolsa de valores americana el índice del valor comercial estaba encima de *NUMERO**
*La tasa promedio acabó la semana con una pérdida neta de *NUMERO**
*La tasa promedio Dow Jones de treinta industrias bajo de *NUMERO* a *NUMERO* puntos*

En las colocaciones anteriores, las ranuras vacías deben ser llenadas con un número (indicado por *NUMERO* en los ejemplos). Más generalmente, las plantillas de frase especifican las categorías gramaticales de las palabras que pueden llenar las ranuras vacías.

Las plantillas de frase son absolutamente representantes de un dominio dado y se repiten muy a menudo de una manera rígida en un sublenguaje dado. Son específicamente útiles para generación de texto.

2.3 Aplicaciones de colocaciones

Como se ha mencionado antes, las colocaciones son útiles en diversas aplicaciones de procesamiento de lenguaje natural. Entre las más significativas tenemos [9]:

- Redacción de texto
- Resolución de ambigüedad sintáctica o análisis sintáctico
- Desambiguación de sentidos de palabras
- Detección y corrección de malapropismos
- Traducción automática
- Reconocimiento de cohesión de texto
- Segmentación en párrafos
- Esteganografía lingüística

A continuación se presenta una breve descripción de cada una de ellas.

Redacción de texto

Una de las aplicaciones principales de las colocaciones es ayudar a cualquier autor a redactar un texto, seleccionando palabras que combinen sintáctica y semánticamente. Hay sistemas que se encargan de llevar a cabo esta tarea automáticamente y se conocen como sistemas generadores de lenguaje.

Resolución de ambigüedad sintáctica o análisis sintáctico

El proceso de la resolución de ambigüedad sintáctica en utilizar conocimiento lingüístico para elegir el árbol sintáctico correcto. Este conocimiento lingüístico se encuentra en un diccionario de colocaciones.

Una idea de los pasos a seguir para este proceso se describe a continuación:

1. Una vez que se tienen los árboles sintácticos posibles de la oración, se extraen todas las relaciones sintácticas (colocaciones) de cada uno de ellos.
2. Se buscan las colocaciones en el diccionario, sumando las frecuencias de todas ellas. Si la colocación no se encuentra, entonces su frecuencia es cero.
3. Se elige el árbol sintáctico que contenga la mayor suma de frecuencias de sus colocaciones.
4. Si el diccionario de colocaciones no tuviera frecuencias entonces se consideraría el valor cero si no existe la colocación y 1 si existe.

Desambiguación de sentidos de palabras

Tomada fuera de contexto, una colocación puede tener diferentes significados, mientras que una colocación adicional puede desambiguar el sentido inmediatamente. Ejemplo, "banco" es: "dinero" si en la base de datos de colocaciones se encuentra "cuenta de banco", es "transfusión" si se encuentra "banco de sangre", es "mueble" si tenemos "sentarse en banco".

5.5.4 Detección y corrección de malapropismos

Malapropismo es un error semántico de reemplazar una palabra real por otra, similar a la deseada en sonido y función sintáctica pero distinta en significado. Ejemplo, centro histérico (queriendo decir centro histórico). Para detectar el malapropismo se propone en [19] basarse en anomalías semánticas en textos originados de los que el malapropismo usualmente destruye el contexto de colocaciones, es decir, la frase dada es sintácticamente correcta pero su colocación no.

La ausencia de tal combinación en la base de datos de colocaciones puede significar un malapropismo en el texto revisado. Para corregir el error es necesario buscar entre las palabras reales las similares a la errónea. Si un candidato restaura la colocación con las otras palabras del contexto, ésta podría ser mostrada al usuario para considerarse.

Traducción automática

Supongamos que tenemos una base de datos de colocaciones en español con una interfaz de opción de traducción. El usuario puede introducir, como consulta, una colocación correcta en un lenguaje diferente al español. Si existen en la base de datos de colocaciones, para cada colocación solicitada, una lista de sus equivalentes en español se mostrará al usuario [21]. Note que en dirección contraria una traducción correcta es generalmente irrealizable.

Reconocimiento de cohesión de texto

El texto "María comió rápidamente, las donas estaban sabrosas", parece consistente para nosotros debido a que donas es comida, esto hace claro que María las comió. Una aplicación puede emular el reconocimiento de cohesión de texto si encuentra la colocación "comer donas" en la base de datos de colocaciones.

Segmentación en párrafos

En [1] se propone un método para segmentación automática de textos en párrafos. La cohesión en la palabra actual es medida por el número de palabras que están dentro de las ligas puramente semánticas y las ligas intracolocación. Una separación de párrafo es colocada cerca mínimo local de profundad definido para la medida de cohesión.

Esteganografía lingüística

Bits de información secreta pueden ser escondidos en un texto que parece inofensivo, seleccionando sinónimos específicos de palabras en un orden previamente acordado, estableciendo la selección del primer sinónimo posible de una palabra por 0, el segundo por 1, etc. Para mantener la cohesión y naturalidad del texto, el sistema elige sólo sinónimos que forman colocaciones de palabras compatibles.

3 Formalismos sintácticos de la lingüística computacional

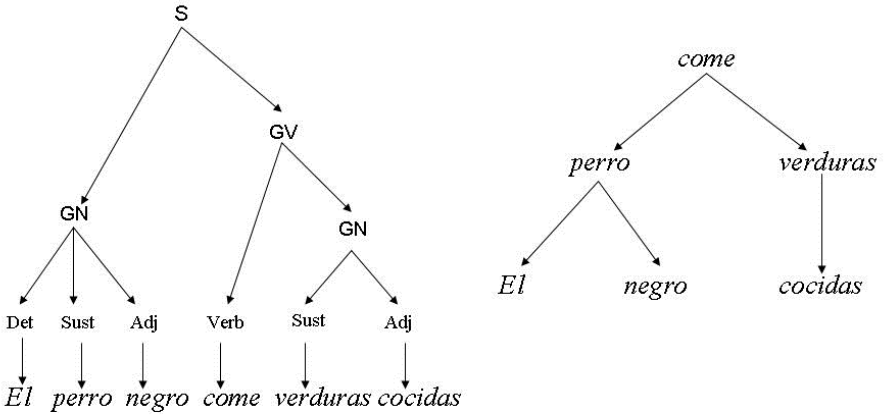
La base para la extracción de colocaciones que se presenta en este artículo es un corpus etiquetado con estructuras sintácticas. Existen dos principales formalismos para representar la estructura sintáctica de una oración: el formalismo de constituyentes (o estructura de frase) cuyo principal representante es la teoría desarrollada por Chomsky en sus diversas variantes; y la tradición estructuralista europea (dependencias) que proviene de Tesniére, con el ejemplo más representativo, la teoría Sentido ⇔ Texto de I. A. Mel'čuk.

Siguiendo el paradigma de Chomsky, se han desarrollado muchos formalismos para la descripción y el análisis sintácticos. El concepto básico de la gramática generativa es simplemente un sistema de reglas que define de una manera formal y precisa un conjunto de secuencias (cadenas a partir de un vocabulario de palabras) que representan las oraciones bien formadas de un lenguaje específico. Las gramáticas bien conocidas en otras ramas de la ciencia de la computación, las expresiones regulares y las gramáticas libres de contexto, son gramáticas generativas también.

Chomsky y sus seguidores desarrollaron y formalizaron una teoría gramatical basada en la noción de generación [22]. El trabajo que se realiza en la gramática generativa descansa en la suposición acerca de la estructura de la oración que está organizada jerárquicamente en frases (y por consiguiente en estructura de frase). Un ejemplo de la segmentación y clasificación que se realiza en este enfoque se presenta en la figura 1A en el árbol de constituyentes para la frase "El perro negro come verduras cocidas".

Un árbol de constituyentes revela la estructura de una expresión en términos de agrupamientos (bloques) de palabras, que consisten de bloques más pequeños, los cuales consisten de bloques aún más pequeños, etc. En un árbol de constituyentes, la mayoría de los nodos representan agrupamientos sintácticos o frases, y no corresponden a las formas de las palabras reales de la oración bajo análisis. Símbolos como S (oración), GN (grupo nominal), GV (grupo verbal), Sust (sustantivo), GP (grupo preposicional), etc. aparecen en los árboles de constituyentes como etiquetas en los nodos, y se supone que estas únicas etiquetas completamente determinan las funciones sintácticas de los nodos correspondientes.

En el enfoque de constituyentes(o estructura de frase), la categorización (la membresía de clase sintáctica) de las unidades sintácticas se especifica como una parte integral de la representación sintáctica, pero no se declaran explícitamente las relaciones entre unidades.



A) Árbol de constituyentes **B) Árbol de dependencias**

Fig. 1. Estructuras sintácticas

Las gramáticas de dependencias se basan en la idea de que la sintaxis es casi totalmente una materia de capacidades de combinación, y en el cumplimiento de los requerimientos de las palabras solas. En el trabajo más influyente en este enfoque, se presenta en [23], el modelo para describir estos fenómenos es semejante a la formación de moléculas, a partir de átomos, en la química. Como átomos, las palabras tienen valencias; están aptas para combinarse con un cierto número y clase de otras palabras, formando piezas más grandes de material lingüístico.

Las valencias de una palabra se rellenan con otras palabras, las cuales realizan dos tipos de funcionamiento: principales (denominadas actuantes) y auxiliares (denominados circunstanciales o modificadores). Las descripciones de valencias de palabras son el dispositivo principal para describir estructuras sintácticas en las gramáticas de dependencias.

La gramática de dependencias supone que hay comúnmente una asimetría entre las palabras de una frase: una palabra es la rectora, algunas otras son sus dependientes. Cada palabra tiene su rectora, excepto la raíz, pero no todas tienen dependientes. Por ejemplo, una palabra es "verduras", la modificadora es "cocidas". La palabra rectora raíz da origen a la construcción total y la determina. Las dependientes se ajustan a las demandas sobre la construcción, impuestas por la rectora. La diferencia entre rectoras y dependientes se refleja por la jerarquía de nodos en el árbol de dependencias.

Las gramáticas de dependencia, como las gramáticas de constituyentes, emplean árboles a fin de describir la estructura de una frase u oración completa. Mientras la gramática de constituyentes asocia los nodos en el árbol con constituyentes mayores o menores y usa los arcos para representar la relación entre una parte y la totalidad, todos los nodos en un árbol de dependencias representan palabras elementales y los arcos denotan las relaciones directas sintagmáticas entre esos elementos (Figura 1B).

Las teorías de constituyentes y las gramáticas de dependencias se han desarrollado en paralelo. Ambas han marcado la forma en la que se concibe la sintaxis en el procesamiento lingüístico de textos. A lo largo de casi cuarenta años, muchos formalismos se han desarrollado dentro de ambos enfoques de una manera muy diferente. A continuación se presenta una un panorama del desarrollo de ambos formalismos.

3.1 Formalismos de constituyentes

Chomsky [24] presentó una versión inicial de la Gramática Generativa Transformacional (GGT), gramática en la cual, la sintaxis se conoce como sintaxis generativa. Una de las características del análisis presentado ahí y en subsecuentes trabajos transformacionales es la inclusión de postulados explícitos formales en las reglas de producción, cuyo único propósito era genera r todas las oraciones gramaticales del lenguaje bajo estudio, es decir, del inglés.

La gramática transformacional inicial influyó, a las teorías posteriores, en el énfasis en la formulación precisa de las hipótesis, característica primordial en el enfoque de constituyentes. Ejemplos de las reglas de producción que se emplean para esa formulación precisa son las siguientes, con las cuales se construyó el árbol de la figura 1A.

O	→	GN GV		ADJ	→	negro	|	cocidas
GN	→	ART SUST ADJ		SUST	→	perro	|	verduras
GN	→	ADJ SUST		V	→	come		
GV	→	V GN		ART	→	el		

La flecha significa que se reescribe como, es decir, el elemento de la izquierda se puede sustituir con el agrupamiento completo de la derecha. Por ejemplo, una oración (O) se puede reescribir como un grupo nominal (GN) seguido de un grupo verbal (GV). Un GN puede reescribirse como un artículo (ART) seguido de un sustantivo (SUST) y un adjetivo (ADJ). Un grupo verbal puede sustituirse con un verbo (V) seguido de un grupo nominal. Todos los elementos que no han sido sustituidos por palabras específicas se denominan no-terminales (GV, O, etc.), los elementos del lenguaje específico se denominan terminales (come, perro, etc.).

Este tipo de reglas corresponde a una gramática independiente del contexto. Esto se debe a que los elementos izquierdos de las reglas solamente contienen un elemento no terminal y por lo tanto no se establece el contexto en el que deben aparecer. Este tipo de gramáticas es el segundo tipo de gramáticas menos restrictivas en la clasificación de Chomsky, que pueden analizarse con un autómata de pila, y para las cuales existen algoritmos de análisis eficientes [25].

Chomsky dio varios argumentos para mostrar que se requería algo más que las solas reglas de estructura de frase para dar una descripción razonable del inglés, y por extensión, de cualquier lenguaje natural, por lo que se requerían las transformaciones, es decir, reglas de tipos más poderosos.

La GGT define oraciones gramaticales de una manera indirecta. Las estructuras aquí denominadas subyacentes o base se generan mediante un sistema de reglas de estructura de frase y después se aplican sucesivamente las reglas transformacionales

para mapear esas estructuras de frase a otras estructuras de frase. Esta sucesión se llama derivación transformacional e involucra una secuencia de estructuras de frase, de una estructura base a una estructura de frase denominada estructura superficial, cuya cadena de palabras corresponde a una oración del lenguaje. Desde este punto de vista, las oraciones del lenguaje son aquellas que pueden derivarse de esta manera.

Una propuesta clave en las gramáticas transformacionales, en todas sus versiones, es que una gramática empíricamente adecuada requiere que las oraciones estén asociadas no con una sola estructura de árbol sino con una secuencia de árboles, cada una relacionada a la siguiente por una transformación. Las transformaciones se aplican de acuerdo a reglas particulares en forma ordenada; en algunos casos las transformaciones son obligatorias.

Otro punto muy importante de la GGT fue el tratamiento del sistema de verbos auxiliares del inglés, el análisis más importante en esta teoría. La GGT inicial se transformó con base a los cambios propuestos en los trabajos de [26] y de [22]. La teoría resultante fue la Teoría Estándar (*Standard Theory*, ST). Entre esos cambios, la ST introdujo el uso de reglas recursivas de estructura de frase para eliminar las transformaciones que combinaban múltiples árboles en uno solo, y la inclusión de características sintácticas, para considerar la subcategorización. Otra aportación fue la adición de una componente semántica interpretativa a la teoría de la gramática transformacional.

Chomsky abandonó algunas ideas de la ST y propuso la Teoría Estándar Ampliada (*The Extended Standard Theory*, EST), una teoría muy reducida en transformaciones, en su lugar se mejoraron otras componentes de la teoría para mantener la capacidad descriptiva. Además de nuevos tipos de reglas semánticas, introdujeron la esquematización de reglas de estructura de frase, y una concepción mejorada del diccionario, incluyendo reglas léxicas. Estas modificaciones se han trasladado a muchos trabajos contemporáneos.

La EST presentó dos modificaciones esenciales:

- El modelo de interpretación semántica debe considerar el conjunto de árboles engendrados por las transformaciones a partir de la estructura profunda.
- El modelo incluye una etapa de inserción léxica antes de la aplicación de las transformaciones. Así que sólo existen dos tipos de reglas: las gramaticales y las de inserción léxica.

Las teorías siguientes a partir de la EST buscaron sobre todo resolver las cuestiones metodológicas debidas a la sobrecapacidad del modelo. [27] y [28] demostraron que el modelo transformacional era equivalente a una gramática sin restricciones.

De hecho, después de varios años de trabajo, esta ba claro que las reglas transformacionales eran muy poderosas y se permitían para toda clase de operaciones que realmente nunca habían si do necesarias en las gramáticas de lenguajes naturales. Por lo que el objetivo de restringir las transformaciones se volvió un tema de investigación muy importante.

Con base en esto Bresnan [29] presenta la Gramática Transformacional Realista que por primera vez proveía un tratamiento convincente de numerosos fenómenos, como la posibilidad de tener forma pasiva en términos léxicos y no en términos trans-

formacionales. Este paso de Bresnan fue se guido por otros investigadores para tratar de eliminar totalmente las transformaciones en la teoría sintáctica.

Otra circunstancia en favor de la eliminación de las transformaciones fue la introducción de la Gramática de Montague [30, 31], ya que al proveer nuevas técnicas para la caracterización de los sentidos, directamente en términos de la estructura superficial, eliminaba la motivación semántica para las transformaciones sintácticas. Con el empleo de métodos de análisis semántico como el de Montague, se podían asignar formalmente distintas estructuras superficiales a distintas pero equivalentes interpretaciones semánticas; de esta manera, se consideraba la semántica sin necesidad de las transformaciones.

Es así como a fines de la década de los setenta y principios de los ochenta surgen los formalismos generativos donde las transformaciones, si existen, tienen un papel menor. Los más notables entre estos son: *Government and Binding* (GB), *Generalized Phrase Structure Grammar* (GPSG), *Lexical-Functional Grammar* (LFG) y *Head-Driven Phrase Structure Grammar* (HPSG), que indican los caminos que han llevado al estado actual en el enfoque de constituyentes.

3.2 Formalismos de dependencias

Mel'cuk [32] explicó que un lenguaje de constituyentes describe muy bien cómo los elementos de una expresión en lenguaje natural combinan con otros elementos para formar unidades más amplias de un orden mayor, y así sucesivamente. Un lenguaje de dependencias, por el contrario, describe cómo los elementos se relacionan con otros elementos, y se concentra en las relaciones entre unidades últimas sintácticas, es decir, entre palabras.

La estructura de un lenguaje también se puede describir mediante árboles de dependencias, los cuales presentan las siguientes características:

- Muestra cuáles elementos se relacionan con otros y en qué forma.
- Revela la estructura de una expresión en términos de ligas jerárquicas entre sus elementos reales, es decir, entre palabras.
- Se indican explícitamente los roles sintácticos, mediante etiquetas especiales.
- Contiene solamente nodos terminales, no se requiere una representación abstracta de agrupamientos.

Con las dependencias se especifican fácilmente los tipos de relaciones sintácticas. Pero la membresía de clase sintáctica (categorización) de unidades de orden más alto (GN, GP, etc.) no se establece directamente dentro de la representación sintáctica misma, así que no hay símbolos no-terminales en representaciones de dependencias.

Una gramática cercana a este enfoque de dependencias es la Gramática Relacional (*Relational Grammar, RG*) [33] que adopta primitivas que son conceptualmente muy cercanas a las nociones relacionales tradicionales de sujeto, objeto directo, y objeto indirecto. Las reglas gramaticales de la RG se formularon en términos relacionales, reemplazando las formulaciones iniciales, basadas en configuraciones de árboles. Por ejemplo, la regla pasiva se establece más en términos de promover el objeto directo al sujeto, que como un re-arreglo estructural de grupos nominales.

Los ejemplos más representativos de este formalismo son: *Dependency Unification Grammar* (DUG), *Word Grammar* (WG) y *Meaning ⇔ Text Theory* (MTT)

4 Extracción del diccionario de colocaciones

En este artículo se presenta un método que obtiene un diccionario de colocaciones en español a partir de un corpus etiquetado manualmente con las estructuras sintácticas. El corpus utilizado es el corpus en español Cast3LB y la extracción del diccionario de colocaciones puede ser descrita en dos pasos:

1. Transformación del corpus de constituyentes a corpus de dependencias
2. Extracción de colocaciones
3. Agregar información estadística

4.1 El corpus en español Cast3LB

Cuenta con cien mil palabras (aproximadamente 3,500 oraciones) creado a partir de dos córporas: el corpus CLiCTALP (75,000 palabras), un corpus balanceado y anotado morfológicamente que contiene un lenguaje literario, periodístico, científico, etc.; y el corpus de la agencia de noticias española EFE (25,000 palabras) correspondiente al año 2000.

El proceso de anotación se llevó a cabo en dos pasos. En el primero, un subconjunto del corpus ha sido seleccionado y anotado dos veces por dos diferentes anotadores. Los resultados de este proceso de doble anotación se han comparado y una topología de desacuerdo en asignación de sentido ha sido establecida. Después de un proceso de análisis y discusión, un manual de anotación ha sido producido, donde los criterios principales a seguir en caso de ambigüedad se han descrito. En el segundo paso, el resto del corpus ha sido anotado siguiendo todas las estrategias de palabras. Los items léxicos anotados son esas palabras con significado léxico, es decir, sustantivos, verbos y adjetivos [34].

4.2 Transformación del corpus de constituyentes a corpus de dependencias

Al igual que la mayoría de las herramientas y recursos existentes, el corpus Cast3LB se orienta a la representación de constituyentes. La extracción de colocaciones (relaciones sintácticas) se hace en base a estructuras sintácticas orientadas a dependencias., es por ello que para llevar a cabo la extracción automática del diccionario de colocaciones el primer paso es transformar el corpus de constituyentes a un corpus de dependencias.

Para esta transformación se llevó a cabo el proceso descrito en [35] y que, en general, se describe a continuación:

1. Extracción de las reglas gramaticales del corpus de constituyentes Cast3LB
2. Determinación de rectores o cabezas de cada regla gramatical mediante el uso de heurísticas

3. Utilizar la información de rectores o cabezas, de forma recursiva, para determinar cuáles reglas y componentes se subirán de nivel en el árbol de dependencias.

Además, como parte de este proceso, se da un tratamiento especial a pronombres y conjunciones mismo que se describe en [36].

4.3 Extracción de colocaciones

Una vez que tenemos el corpus de dependencias aplicamos los siguientes pasos para extraer las colocaciones:

1. Recorremos el árbol de dependencias en profundidad de izquierda a derecha, comenzando de la raíz.
2. Por cada nodo hijo del nodo visitado, se extrae el nodo padre, el nodo hijo y la relación de dependencia entre ellos. Si el nodo hijo es una preposición entonces éste se considera como la relación de dependencia y el nodo hijo de la preposición se considera el nodo hijo de la colocación.

No se consideran las colocaciones donde existen determinantes (artículos) debido a que, como mencionamos anteriormente, las colocaciones son pares de palabras de contenido con su relación sintáctica.

Para ilustrar un ejemplo consideramos la siguiente oración: "Los policías velarán por la seguridad de los líderes" (el árbol de dependencias extraído de esta oración se muestra en la figura 2).

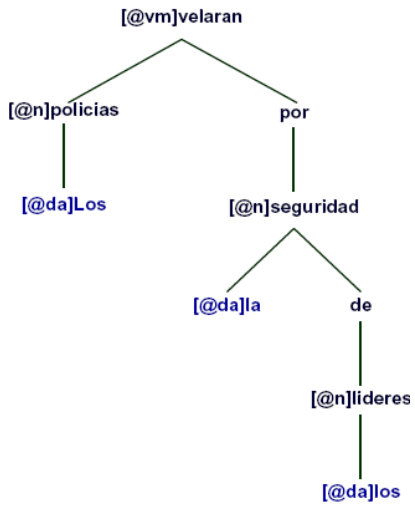


Fig. 2. Árbol sintáctico de dependencias para la oración "Los policías velarán por la seguridad de los líderes"

Recorriendo el árbol (de la figura 2), visitamos el primer nodo que sería la raíz y encontramos que la primer colocación a extraer es *velarán SUST policías*, donde *vela-*

rán es el nodo padre, *policías* es el nodo hijo y *SUST* es la relación de dependencia entre ellos. Las colocaciones extraídas automáticamente de la oración "Los policías velarán por la seguridad de los líderes" son las siguientes y se muestran encerradas en óvalos en la figura 3:

seguridad de líder

velar SUST policía

velar por seguridad

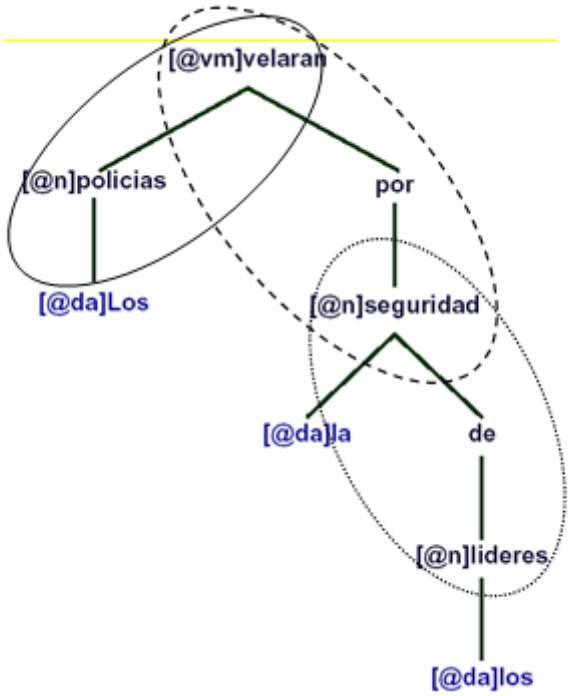


Fig. 3. Colocaciones extraídas del árbol de dependencias de la oración "Los policías velarán por la seguridad de los líderes".

4.4 Agregar información estadística

Por último, para las frecuencias de las colocaciones del diccionario se llevaron a cabo los siguientes pasos:

1. Se ordenaron las colocaciones obtenidas
2. Se cuentan las frecuencias de las colocaciones
3. Se eliminan las colocaciones repetidas

Esta información fue agregada ya que se considera importante para algunas aplicaciones de procesamiento de lenguaje natural, específicamente, para resolver la ambigüedad sintáctica.

5 Resultados

El diccionario de colocaciones en español extraído automáticamente del corpus Cast3LB consta de 40,121 colocaciones únicas. Cada una de esas colocaciones está formada por la palabra rectora (o nodo padre), la palabra dependiente (nodo hijo) y la relación sintáctica entre ambas, así como al frecuencia de dicha colocación.

En el diccionario se encuentran tres tipos de colocaciones: 13,048 oraciones nominales rígidas, 423 plantillas de frase y 26,650 relaciones predicativas.

Con respecto a las colocaciones de tipo relaciones predicativas, en el diccionario se clasifican de acuerdo a los siguientes tipos de relación sintáctica entre palabras: sustantivo, adjetivo, verbo, adverbio, pronombre, coordinante, negación y preposición.

La tabla siguiente muestra el número de colocaciones que contiene el diccionario con respecto al tipo de colocación y la relación sintáctica correspondiente, así como un ejemplo para cada caso.

Tabla 3. Colocaciones extraídas automáticamente de acuerdo al tipo de colococación

Tipo de colocación	Relación sintáctica de la colocación	Número de colocaciones obtenidas	Ejemplo
Relaciones predicativas	Sustantivo	8,821	1 aceptar SUST suerte
	Adjetivo	5,802	1 águila ADJ real
	Verbo	5,370	1 acuerdo VERB establecer
	Adverbio	2,692	1 acabar ADV bien
	Pronombre	2,103	1 acabar PRON ese
	Coordinante	1,364	1 prestar COORD si
	Negación	293	1 aprovechar NEG no
	Preposición	205	1 arrinconar PREP contra
Oraciones nominales rígidas	----	13,048	1 ansia de revolución
Plantillas de frase	----	423	1 alcanzar CIF *número* Antes_de FECH *fecha*

Con relación a la información estadística del diccionario, 36,501 colocaciones tienen frecuencia de 1, es decir, sólo aparecen una vez en el corpus; 3,524 colocaciones tienen frecuencia de 2 a 10; 94 tienen frecuencia de 11 a 100, mientras que sólo dos colocaciones contienen una frecuencia mayor a 100.

5.1 Evaluación

Para evaluar las colocaciones obtenidas automáticamente, se seleccionaron dos muestras. Cada una consiste de 17 oraciones seleccionadas aleatoriamente de todas las oraciones del corpus Cast3LB.

El sistema extrae automáticamente las colocaciones de cada oración en ambas muestras. Para poder compararlas, un experto extrajo manualmente las colocaciones de las oraciones de cada muestra.

La tabla 4 muestra los resultados de las colocaciones extraídas manualmente, automáticamente, colocaciones que coinciden, la precisión y el *recall* por cada oración de la primera muestra seleccionada.

Tabla 4. Resultados obtenidos para la primer muestra

Oración	Colocaciones extraídas manualmente	Colocaciones extraídas automáticamente	Colocaciones que coinciden	Precisión	*Recall*
1	15	15	13	86.7	86.7
2	28	30	21	70.0	75.0
3	9	9	8	88.9	88.9
4	12	12	11	91.7	91.7
5	11	11	10	90.9	90.9
6	6	6	6	100	100
7	9	9	5	55.6	55.6
8	15	15	15	100	100
9	11	11	11	100	100
10	17	17	16	94.1	94.1
11	9	6	6	100	66.7
12	13	10	10	100	76.9
13	7	7	7	100	100
14	14	13	12	92.3	85.7
15	12	12	12	100	100
16	15	15	12	80.0	80.0
17	8	8	8	100	100
			Promedio:	**91.2**	**87.8**

La tabla 5 contiene los mismos resultados pero de la segunda muestra seleccionada. Por precisión se entiende el porcentaje de las colocaciones extraídas automáticamente que también fueron extraídas manualmente, mientras que por *recall* se entiende el porcentaje de las colocaciones extraídas manualmente que también fueron extraídas automáticamente.

El promedio de la precisión de la unión de las dos muestras es de 89.7, con una desviación estándar de 10.2, y el promedio de *recall* es de 88.7 con una desviación estándar de 11.3. La desviación estándar entre los dos promedios de las dos muestras es 0.5% para la precisión y 0.9% para el *recall*. Entonces, para otras muestras extraí-

das del mismo corpus, el porcentaje promedio de colocaciones correctas será similar a los valores obtenidos para las muestras seleccionadas.

El diccionario extraído del corpus Cast3LB cuenta con 40,121 colocaciones. Extrapolando los resultados, inferimos que 89.7% ± 0.5% de ellas son correctas (89.7% ± 10.2% en cada oración específica), y que el diccionario extraído contiene 88.7% ± 0.9% de las colocaciones contenidas realmente en el corpus (88.7% ± 11.3% de cada oración específica).

Tabla 5. Resultados obtenidos para la segunda muestra

Oración	Colocaciones extraídas manualmente	Colocaciones extraídas automáticamente	Colocaciones que coinciden	Precisión	*Recall*
1	3	3	3	100	100
2	18	20	18	90.0	100
3	8	8	6	75.0	75
4	10	10	10	100	100
5	9	9	7	77.8	77.8
6	8	8	8	100	100
7	5	5	5	100	100
8	13	13	11	84.6	84.6
9	16	17	15	88.2	93.8
10	18	18	18	100	100
11	12	12	11	91.7	91.7
12	12	12	11	91.7	91.7
13	11	11	9	81.8	81.8
14	8	10	6	60.0	75.0
15	8	8	7	87.5	87.5
16	13	13	10	76.9	76.9
17	33	31	29	93.5	87.9
			Promedio:	**88.2**	**89.6**

6 Conclusiones y trabajo futuro

El proceso de etiquetar corpus con estructuras sintácticas es una tarea que se sigue llevando acabo. Estos recursos son la base necesaria para extraer colocaciones automáticamente y no enfocar esfuerzos en la generación manual de diccionarios de colocaciones.

En este artículo se llevó a cabo la extracción automática de un diccionario de colocaciones en español basado el corpus etiquetado en español Cast3LB. El diccionario obtenido contiene más de 40,000 colocaciones con más del 89% de precisión.

Como trabajo futuro se propone mejorar el método de extracción de colocaciones, aquí presentado, para obtener mejores resultados, un ejemplo sería el uso del "que" como preposición en algunos grupos verbales (ejemplo: "tienen que tener", donde "que" sería la relación sintáctica entre los dos verbos).

Referencias

1. Bolshakov, I.A., Gelbukh, A.: A very large database of collocations and semantic links. Lecture Notes in Computer Science, N 1959, Springer, pp. 103–114 (2001).
2. Bolshakov, I.A., Gelbukh, A.: Enseñando idiomas extranjeros con una base de colocaciones. In: Caridad Anías Calderón, En: La telemática y su aplicación en la educación a distancia y en la informatización de la sociedad. Editorial "Félix Varela", Cuba, Tomo II, p. 632–638 (2002).
3. Nakhimovsky, A. D., Leed, R. L.: Lexical functions and language learning. Slavic and East European Journal (1979).
4. Allerton, D. J.: Three or four levels of co-occurrence relations. Lingua, 63, pp. 17-40 (1984).
5. Cruse, D. A.: Lexical Semantics. Cambridge University Press. Cambridge, United Kingdom (1986).
6. Mel'cuk, I. A.: Meaning-Text models: A recent trend in Soviet linguistics. In: Annual Review of Anthropology 10, pp. 27-62 (1981).
7. Gelbukh, A., Kolesnikova, O: Semantic Analysis of Verbal Collocations with Lexical Functions. Springer, XI + 146 pp. (2013).
8. Benson, M.: Collocations and general-purpose dictionaries. In: International Journal of Lexicography, 3(1), pp. 23-35 (1990).
9. Bolshakov, I. A.: Getting One's First Million… Collocations. In: CICLing 2004, LNCS 2945 pp. 229-242 (2004).
10. Dellenbaugh, D., Dellenbaugh, B.: Small Boat Sailing, a Complete Guide. Sports Illustrated Winner's Circle Books (1990)
11. Smadja, F. A.: Retrieving Collocations from Text: Xtract. In: Computational Linguistics 19.1: pp. 143-176 (1993)
12. Cowie, A. P.: The treatment of collocations and idioms in learner's dictionaries. In: Applied Linguistics, 2(3), pp. 223-235 (1981)
13. Benson, M.: The collocational dictionary and the advanced learner. In: Learner's Dictionaries: State of the Art, edited by M. Tickoo, 84-93. SEAMEO, (1989).
14. Sidorov, G.: Métodos de análisis de la combinabilidad de palabras en ruso (en ruso). In: Taal en cultur, Maastricht, Holanda y Moscú, Rusia, pp. 294-302 (1999).
15. Benson, M., Benson, E., Ilson, R.: The BBI combinatory dictionary of English: a guide to word combinations. John Benjamins Publishing Company, Philadelphia (1986).
16. Halliday, M. A. K.: Lexis as a linguistic level. In: In Memory of J. R. Firth, edited by C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins, Longmans Linguistics Library, pp. 148-162, (1966).
17. Choueka, Y.: Looking for needles in a haystack. In: Proceedings, RIAO Conference on User-Oriented Context Based Text and Image Handling, pp. 609-623. Cambridge, MA (1988).
18. Amsler, B.: Research towards the development of a lexical knowledge base for natural language processing. In: Proceedings, 1989 SIGIR Conference. Cambridge, MA (1989).
19. Bolshakov, I. A., Gelbukh, A.: On detection of Malapropisms by Multistage Collocation Testing. In: NLDB-2003, 8th Int. Conf. on Application on Natural Lengauge to Information Systems. Bonner Köllen Verlag, pp. 28-41 (2003).
20. Bolshakov, I. A., Gelbukh, A.: Text segmentation to Paragraphs based on Local Text Cohesion. In: V. Matousek et al. (Eds). Text Speech and Dialogue. Proc. 4th Intern. Conf. TSD-2001. Lecture Notes in Artificial Intelligence, 2166, Springer, pp. 158-166 (2001).

21. Bolshakov, I.A., Gelbukh, A.: A Large Database of Collocations and Semantic References: Interlingual Applications. International Journal of Translation, Vol. 13, No.1–2, pp. 167–187 (2001).
22. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, Cambridge, MA (1965).
23. Tesniere, L.: Elements de syntaxe structural. Paris: Klincksiek. (German: Tesniere, L. (1980): Grundzüge der strukturalen Syntax. Stuttgart: Klett-Cotta.) (1959).
24. Chomsky, N.: Syntactic Structures. The Hague: Mouton & Co (1957)
25. Aho, A. V., Sethi, R., Ullman, J. D.: Compilers. Principles, Techniques and Tools. Addison Wesley Publishing Company, (1986).
26. Katz, J. J., Postal, P. M.: An Integrated Theory of Linguistic Descriptions. Cambridge, Mass. MIT Press, (1964).
27. Salomaa, A.: The generative power of transformational grammars of Ginsburg and Partee. Information and Control, 18, pp. 227-232 (1971).
28. Peters, P. S., Ritchie, R. W.: On the generative power of transformational grammars. Information Science, 6, pp. 49 - 83 (1973).
29. Bresnan, J. A.: Realistic transformational Grammar. In M. Halle, J. Bresnan and G. A. Miller (eds.), Linguistic Theory and Psychological Reality. Cambridge, Mass. MIT Press (1978).
30. Montague, R.: Universal Grammar. Theoria 36, pp. 373- 398 (1970).
31. Montague, R.: Universal Grammar. En: Richard Thomason (eds.), Formal Philosophy. New Haven: Yale University Press (1974).
32. Mel'cuk, I. A.: Dependency Syntax. In: P. T. Roberge (ed.) Studies in Dependency Syntax. Ann Arbor: Karoma pp. 23-90 (1979)
33. Perlmutter, D. N.: Studies in Relational Grammar I. Chicago: University of Chicago Press (1983).
34. Navarro, B., Civit, M., Martí, M.A., Marcos, R., Fernández, B.: Syntactic, Semantic and Programatic Annotation in Cast3LB. In: Shallow Processing of Large Corpora (SProLaC), a Workshop on Corpus linguistic, Lancaster, UK (2003).
35. Gelbukh, A., Torres, S., Calvo, H.: Transforming a constituency Treebank into a dependency Treebank. Procesamiento del lenguaje natural, vol. 35, pp, 145–152, España (2005).
36. Gelbukh, A., Torres, S.: Tratamiento de ciertos pronombres y conjunciones en la transformación de un corpus de constituyentes a un corpus de dependencias. In: Avances en la Ciencia de la Computación.. (Eds.) Arturo Hernández Aguirre, Jose Luis Zechinelli Martini, pp. 293-298, México (2006).