



INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
Laboratorio de Lenguaje Natural



**Aprendizaje supervisado de colocaciones
para la resolución de la ambigüedad
sintáctica**

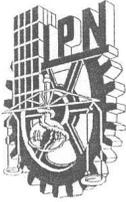
T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRA EN CIENCIAS DE LA COMPUTACIÓN
PRESENTA

SULEMA TORRES RAMOS

Director:
Dr. Alexander Gelbukh

México, D.F.
Junio, 2006



INSTITUTO POLITECNICO NACIONAL
COORDINACION GENERAL DE POSGRADO E INVESTIGACION

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D. F. Siendo las 16:00 horas del día 12 del mes de Junio de 2006 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis de grado titulada:

**" APRENDIZAJE SUPERVISADO DE COLOCACIONES PARA LA
RESOLUCIÓN DE LA AMBIGÜEDAD SINTÁCTICA"**

Presentada por la alumna:

TORRES

Apellido paterno

RAMOS

Materno

SULEMA

nombre(s)

Con registro:

B	0	4	1	2	6	9
---	---	---	---	---	---	---

aspirante al grado de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron **SU APROBACIÓN DE LA TESIS**, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Presidente

Dr. Igor Bolshakov

Secretario

Dr. Sergio Suárez Guerra

Primer vocal

Dr. Alexandre Guelboukh Kahn

Segundo vocal

Dr. Grigori Sidorov

Tercer vocal

Dra. Sofia Natalia Galicia Haro

Suplente

Alejandro Botello Castillo

EL PRESIDENTE DEL COLEGIO

DR. HUGO CÉSAR COY



INSTITUTO POLITECNICO NACIONAL

COORDINACIÓN GENERAL DE POSGRADO E INVESTIGACIÓN
EN COMPUTACIÓN

DIRECCIÓN



INSTITUTO POLITECNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESION DE DERECHOS

En la Ciudad de México el día 12 del mes de Junio del año 2006, el que suscribe *Sulema Torres Ramos* alumna del *Programa de Maestría en Ciencias de Computación* con número de registro *B041269*, adscrita al *Centro de Investigación en Computación*, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del *Dr Alexander Gelbukh Kahn* del trabajo intitulado *Aprendizaje supervisado de colocaciones para la resolución de la ambigüedad sintáctica* cede los derechos, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección *sulema7@hotmail.com*. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.


Sulema Torres Ramos

Nombre y firma

Abstract

Natural languages are full of collocations, which are recurrent and arbitrary combinations of words that co-occur more often than expected by chance.

Collocations are useful in a variety of applications of natural language processing. One of these applications is solving syntactic ambiguity, which is one of the hardest challenges to overcome in natural language processing systems.

The idea of solving this kind of ambiguity by using collocations is based in the following principle: when there are two or more syntactic trees of the same sentence, all syntactic relations —collocations— are extracted from each tree, then we search for these syntactic relations in a collocations dictionary, and lastly the syntactic tree which contains the largest number of found collocations is chosen.

The main contribution presented in this thesis is the extraction of a collocations dictionary.

Resumen

Los lenguajes naturales están llenos de colocaciones, combinaciones recurrentes y arbitrarias de palabras que co-ocurren frecuentemente y no por casualidad.

Las colocaciones son útiles en diferentes aplicaciones de procesamiento de lenguaje natural. Una de ellas resolver la ambigüedad sintáctica que es uno de los problemas más difíciles que se presentan en sistemas de procesamiento de lenguaje natural.

La idea de resolver este tipo de ambigüedad utilizando colocaciones se basa en el siguiente principio: cuando se tienen dos o más árboles sintácticos de una misma oración se extraen todas las relaciones sintácticas (colocaciones) de cada árbol, éstas se buscan en un diccionario de colocaciones y al final se elige el árbol sintáctico que contenga el mayor número de colocaciones encontradas.

La aportación principal de esta tesis consiste en la extracción de un diccionario de colocaciones.

Contenido general de la tesis

ABSTRACT	II
RESUMEN	III
CONTENIDO GENERAL DE LA TESIS	IV
ÍNDICE DETALLADO DE LA TESIS	V
LISTA DE FIGURAS.....	VIII
LISTA DE TABLAS.....	IX
CAPÍTULO 1. INTRODUCCIÓN	2
CAPÍTULO 2. LA ESTRUCTURA SINTÁCTICA EN EL LENGUAJE NATURAL Y EL PROBLEMA DE LA AMBIGÜEDAD SINTÁCTICA	9
CAPÍTULO 3. LOS FORMALISMOS SINTÁCTICOS DE CONSTITUYENTES Y DE DEPENDENCIAS	19
CAPÍTULO 4. TRANSFORMACIÓN DEL CORPUS DE CONSTITUYENTES A UN CORPUS DE DEPENDENCIAS	44
CAPÍTULO 5. EXTRACCIÓN DEL DICCIONARIO DE COLOCACIONES	60
CAPÍTULO 6. EVALUACIÓN DEL PARSER DE DEPENDENCIAS DILUCT.....	71
CAPÍTULO 7. CONCLUSIONES Y TRABAJO FUTURO.....	80
GLOSARIO	85
REFERENCIAS	93
ÍNDICE DE TÉRMINOS	103
ANEXOS	105

Índice detallado de la tesis

ABSTRACT	II
RESUMEN	III
CONTENIDO GENERAL DE LA TESIS	IV
ÍNDICE DETALLADO DE LA TESIS	V
LISTA DE FIGURAS.....	VIII
LISTA DE TABLAS.....	IX
CAPÍTULO 1. INTRODUCCIÓN.....	2
1.1 UBICACIÓN Y ALCANCE.....	2
1.2 OBJETIVOS	3
1.2.1 <i>Objetivo general</i>	3
1.2.2 <i>Objetivos específicos</i>	3
1.3 RELEVANCIA E IMPORTANCIA.....	4
1.4 NOVEDAD CIENTÍFICA.....	4
1.5 APORTACIONES PRINCIPALES.....	5
1.5.1 <i>Aportaciones teóricas</i>	5
1.5.2 <i>Productos obtenidos</i>	6
1.6 ORGANIZACIÓN DE LA TESIS	6
CAPÍTULO 2. LA ESTRUCTURA SINTÁCTICA EN EL LENGUAJE NATURAL Y EL PROBLEMA DE LA AMBIGÜEDAD SINTÁCTICA	9
2.1 INTRODUCCIÓN.....	9
2.2 LENGUAJE.....	9
2.2.1 <i>Lenguaje Natural</i>	10
2.2.2 <i>Lenguaje Formal</i>	11
2.3 PROCESAMIENTO DEL LENGUAJE NATURAL	11
2.4 NIVELES DEL LENGUAJE	12
2.5 AMBIGÜEDAD	13
2.5.1 <i>Tipos de ambigüedad</i>	13
2.6 AMBIGÜEDAD SINTÁCTICA	14
<i>Patrones de manejo</i>	15
<i>Reglas ponderadas</i>	15
<i>Proximidad semántica</i>	15
2.7 USO DE COLOCACIONES PARA RESOLVER AMBIGÜEDAD SINTÁCTICA	16
2.8 CONCLUSIONES.....	17
CAPÍTULO 3. LOS FORMALISMOS SINTÁCTICOS DE CONSTITUYENTES Y DE DEPENDENCIAS	19
3.1 INTRODUCCIÓN.....	19
3.2 FORMALISMOS DE LA LINGÜÍSTICA COMPUTACIONAL.....	20

3.2.1 Formalismos de constituyentes	22
3.2.1.1 Government and Binding (GB)	26
3.2.1.2 Generalized Phrase Structure Grammar (GPSG)	27
3.2.1.3 Lexical-Functional Grammar (LFG)	28
3.2.1.4 Head-Driven Phrase Structure Grammar (HSPG).....	30
3.2.2 Formalismos de dependencias	32
3.2.2.1 Dependency Unification Grammar (DUG).....	33
3.2.2.2 Word Grammar (WG).....	34
3.2.2.3 Meaning \Leftrightarrow Text Theory (MMT).....	35
3.2.3 Formalismos mixtos	38
3.3 COMPARACIÓN DE LOS FORMALISMOS SINTÁCTICOS	40
3.4 CONCLUSIONES.....	42
CAPÍTULO 4. TRANSFORMACIÓN DEL CORPUS DE CONSTITUYENTES A UN CORPUS DE DEPENDENCIAS	44
4.1 INTRODUCCIÓN	44
4.2 EL CORPUS EN ESPAÑOL CAST3LB	45
4.3 EXTRACCIÓN DE LA GRAMÁTICA DE CONSTITUYENTES	45
4.4 HEURÍSTICAS PARA LA DETERMINACIÓN DE LOS JEFES DE FRASES EN LAS REGLAS DE GRAMÁTICA ..	48
4.5 EVALUACIÓN DE LAS HEURÍSTICAS Y CORRECCIÓN DE LA GRAMÁTICA	51
4.6 ALGORITMO PARA LA TRANSFORMACIÓN DE UN ÁRBOL DE CONSTITUYENTES EN UNO DE DEPENDENCIAS	51
4.6.1 El caso de los coordinantes	54
4.7 IMPLEMENTACIÓN	56
4.8 METODOLOGÍA DE EVALUACIÓN.....	57
4.9 RESULTADOS OBTENIDOS	57
4.10 CONCLUSIONES.....	58
CAPÍTULO 5. EXTRACCIÓN DEL DICCIONARIO DE COLOCACIONES	60
5.1 INTRODUCCIÓN	60
5.2 ¿QUÉ SON LAS COLOCACIONES?	60
5.3 PROPIEDADES DE LAS COLOCACIONES	61
5.3.1 Las colocaciones son arbitrarias	61
5.3.2 Las colocaciones son dependientes del dominio.....	62
5.3.3 Las colocaciones son recurrentes	62
5.3.4 Las colocaciones son conjuntos de cohesión léxica.....	62
5.4 TIPOS DE COLOCACIONES.....	63
5.4.1 Relaciones predicativas	64
5.4.2 Oraciones nominales rígidas	64
5.4.3 Plantillas de frase	65
5.5 APLICACIONES DE COLOCACIONES	65
5.6 ALGORITMO PARA LA EXTRACCIÓN DEL DICCIONARIO DE COLOCACIONES DEL CORPUS	66
5.7 IMPLEMENTACIÓN	67
5.8 RESULTADOS OBTENIDOS	68
5.9 CONCLUSIONES.....	69
CAPÍTULO 6. EVALUACIÓN DEL PARSER DE DEPENDENCIAS DILUCT.....	71
6.1 INTRODUCCIÓN	71
6.2 FUNCIONAMIENTO DEL PARSER DILUCT	72
6.2.1 Pre-procesamiento.....	72
6.2.2 Reglas.....	73
6.2.3 Asignación de frase preposicional.....	76
6.2.4 Heurísticas.....	76
6.2.5 Selección de la raíz	77
6.3 ALGORITMO PARA LA EVALUACIÓN DEL PARSER DILUCT CON EL CORPUS CAST3LB	77

6.4 RESULTADOS OBTENIDOS	78
6.5 CONCLUSIONES.....	78
CAPÍTULO 7. CONCLUSIONES Y TRABAJO FUTURO.....	80
7.1 DISCUSIÓN.....	80
7.2 CONCLUSIONES.....	81
7.3 APORTACIONES.....	82
7.3.1 <i>Aportaciones al conocimiento</i>	82
7.3.2 <i>Aportaciones técnicas</i>	82
7.4 PUBLICACIONES GENERADAS.....	83
7.5 TRABAJO FUTURO	83
GLOSARIO	85
REFERENCIAS	93
ÍNDICE DE TÉRMINOS	103
ANEXOS	105
ANEXO 1: MUESTRAS DEL CORPUS CAST3LB	105
ANEXO 2: LAS REGLAS PRINCIPALES DE LA GRAMÁTICA EXTRAÍDA	108
ANEXO 3: MUESTRAS DE LOS ÁRBOLES DE DEPENDENCIAS CONSTRUIDOS	109
ANEXO 4: MUESTRAS DEL DICCIONARIO EXTRAÍDO	112

Lista de figuras

Figura 1. Árboles sintácticos extraídos de la oración “Juan vio a un hombre con un telescopio”	14
Figura 2. Estructuras sintácticas	21
Figura 3. Niveles de Representación en la MTT.	36
Figura 4. Ejemplo de estructura de dependencias en la MTT	37
Figura 5. Oración con etiquetas del corpus Cast3LB (“Las reservas de oro y divisas de Rusia subieron 800 millones de dolares”).	46
Figura 6. Oración simplificada del corpus Cast3LB (“Las reservas de oro y divisas de Rusia subieron 800 millones de dolares”).	47
Figura 7. Patrones a extraer de la oración “Las reservas de oro y divisas de Rusia subieron 800 millones de dólares”	48
Figura 8. Patrones extraídos de la oración “Las reservas de oro y divisas de Rusia subieron 800 millones de dólares”	48
Figura 9. Reglas más frecuentes extraídas del corpus Cast3LB.	52
Figura 10. Reglas marcadas automáticamente no coinciden con las marcadas manualmente.	52
Figura 11. Árbol de constituyentes. ‘Las reservas de oro y divisas de Rusia subieron 800 millones de dolares’	53
Figura 12. Nodos cabeza colocados en la posición del nodo padre del patrón.	53
Figura 13. Primer árbol de dependencias resultante con etiquetas.	54
Figura 14. Segundo de dependencias resultante con etiquetas.	55
Figura 15. Primer árbol de dependencias sin etiquetas.	55
Figura 16. Segundo árbol de dependencias sin etiquetas.	55
Figura 17. Test llenar-el-espacio, de [Benson, 1990].	63
Figura 18. Colocaciones a extraer del primer árbol de dependencias de la oración “Las reservas de oro y divisas de Rusia subieron 800 millones de dólares”	67
Figura 19. Colocaciones extraídas automáticamente de la oración “Las reservas de oro y divisas de Rusia subieron 800 millones de dólares”	67

Lista de tablas

Tabla 1. Descripción de la implementación de los módulos principales del algoritmo de transformación.	56
Tabla 2. Comparaciones lingüísticas cruzadas de colocaciones.	61
Tabla 3. Descripción de la implementación de los módulos principales del algoritmo de extracción.	68

CAPÍTULO

1

1

Introducción

Capítulo 1. Introducción

1.1 Ubicación y alcance

La información es el recurso más importante que poseemos los seres humanos. Gran parte de esta información se comunica, almacena y maneja en forma de lenguaje natural (español, inglés, ruso, etc.). En la actualidad, podemos obtener grandes volúmenes de información en forma escrita, ya sea de manera impresa o electrónica.

Las computadoras son una herramienta indispensable para el procesamiento de la información plasmada en los textos, ya que son más capaces de manejar grandes volúmenes de datos que las personas. Sin embargo, una computadora no puede hacer todo lo que las personas normalmente hacemos con el texto, por ejemplo, responder preguntas basándose en la información proporcionada, o, hacer inferencias lógicas sobre su contenido, o, elaborar un resumen de esta información.

Por lo anterior, el Procesamiento de Lenguaje Natural (PLN) tiene por objetivo habilitar a las computadoras para que entiendan el texto, procesándolo por su sentido. Para llevar a cabo esta tarea, un sistema de PLN necesita conocer sobre la estructura del lenguaje, la cual se analiza normalmente en 4 niveles: morfológico, sintáctico, semántico y pragmático. En el nivel morfológico se estudia cómo se construyen las palabras; en el sintáctico, cómo combinar las palabras para formar oraciones; en el semántico, el significado de las palabras, y por último, en el pragmático se estudia cómo el contexto afecta a la interpretación de las oraciones. Nuestra investigación se centra en el nivel sintáctico.

Todos los niveles anteriores de la estructura del lenguaje tienen un problema: la ambigüedad. Ésta se presenta cuando pueden admitirse distintas interpretaciones de un

texto, oración o palabra. Resolver la ambigüedad es uno de los principales objetivos del PLN. Existen varios tipos de ambigüedad: sintáctica (estructural), léxica y semántica.

En el presente trabajo nos centramos en el problema de la ambigüedad sintáctica, que ocurre cuando existe más de una forma de interpretar la estructura de una oración. Ejemplo: la oración “Veo un hombre con lentes” se puede interpretar de dos formas, utilizo los lentes para ver al hombre o el hombre que veo tiene lentes.

Una forma de resolver la ambigüedad sintáctica es utilizando colocaciones [\[McKeown & Radev, 98\]](#). Una colocación es la relación entre dos palabras o un grupo de palabras que frecuentemente se usan de manera conjunta formando una expresión común. Algunos ejemplos de colocaciones son *sistema político*, *seguro de vida*, *núcleo familiar*, etc. **Esta tesis está enfocada al aprendizaje supervisado de éstas colocaciones.**

1.2 Objetivos

1.2.1 Objetivo general

Desarrollar los métodos y recursos para el aprendizaje supervisado de colocaciones, es decir, la extracción supervisada de un diccionario de colocaciones en español a partir de un corpus de texto etiquetado con las estructuras sintácticas en el formalismo de constituyentes, y aplicar los recursos obtenidos a la evaluación de un analizador sintáctico de dependencias.

1.2.2 Objetivos específicos

1. Extraer la gramática del corpus Cast3LB.
2. Aplicar reglas de heurística para determinar los jefes de frase de la gramática.
3. Aplicar un algoritmo para transformar los árboles de constituyentes en árboles de dependencias.
4. Extraer el diccionario de colocaciones del corpus.
5. Evaluar los resultados obtenidos.

1.3 Relevancia e importancia

La importancia de un diccionario de colocaciones se puede ver reflejada en muchas áreas y aplicaciones. Una de ellas es la desambiguación, tanto sintáctica como semántica. La ambigüedad sintáctica (estructural) es uno de los problemas principales a resolver en sistemas de procesamiento de lenguaje natural.

Otra tarea importante es la traducción automática, debido a que las colocaciones no se pueden caracterizar en base a regularidades sintácticas y semánticas, no pueden ser traducidas sobre una base palabra-por-palabra. Es necesario utilizar técnicas estadísticas enfocadas a alinear corpus bilingües para identificar traducción de colocaciones y así crear semi-automáticamente diccionarios bilingües de colocaciones.

Finalmente, las colocaciones han sido usadas como parte de los sistemas generadores de lenguaje. Estos sistemas utilizan diccionarios de colocaciones y frases de palabra durante el proceso de selección de una palabra.

1.4 Novedad científica

Actualmente se nota la tendencia de uso cada vez mayor de gramáticas de dependencias. Sin embargo, los recursos léxicos existentes, especialmente los corpus con etiquetado sintáctico manual, en su mayoría son orientados al formalismo de constituyentes –en parte por la inercia de las escuelas tradicionales y en parte porque resultan de proyectos de largo plazo cuyo desarrollo se empezó hace varios años.

La creación de corpus orientados a constituyentes es una tarea que continúa hoy en día, de ahí surge la necesidad y relevancia de la obtención automática o semiautomática de los recursos sintácticos para el formalismo de dependencias utilizando los recursos existentes en el formalismo de constituyentes.

Tener un método para transformar automáticamente corpus de constituyentes a dependencias tiene ventajas debido a que cada corpus se limita al trabajo elaborado por un grupo de personas específicas que utilizan ciertos criterios, unir dos corpus de

constituyentes creados por diferentes personas es difícil, mientras que un algoritmo de transformación puede tomar varios corpus de constituyentes para crear nuevos corpus.

Este trabajo propone las heurísticas para la conversión de un corpus de constituyentes a un corpus de dependencias. El corpus obtenido permite la extracción de una gramática de dependencias y la composición de una base de datos de colocaciones sintácticas, así como la evaluación de los analizadores sintácticos de dependencias. Las heurísticas de la conversión son novedosas, así como ciertos detalles del proceso de la extracción de la gramática, de la extracción de las colocaciones, y de la evaluación de los analizadores sintácticos.

1.5 Aportaciones principales

1.5.1 Aportaciones teóricas

- Desarrollo, implementación y evaluación de las heurísticas para la **conversión** de un corpus (*treebank*) con estructuras sintácticas etiquetadas con de constituyentes al corpus (*treebank*) de dependencias.
- Desarrollo, implementación y evaluación de las heurísticas para la extracción, a partir de un corpus de constituyentes, de una **gramática** de gran escala capaz de construir los árboles sintácticos de dependencia.
- Desarrollo, implementación y evaluación de las heurísticas para la extracción de **colocaciones** sintácticas de un corpus etiquetado con constituyentes, incluido el tratamiento adecuado de preposiciones y conjunciones.
- Desarrollo, implementación y aplicación de algoritmo de **evaluación** de un analizador sintáctico de dependencias a través de un corpus existente etiquetado con las estructuras sintácticas de constituyentes.

1.5.2 Productos obtenidos

- **Corpus** de español etiquetado con las **dependencias** sintácticas, obtenido automáticamente a través de conversión de un corpus de constituyentes.
- **Gramática** de español a gran escala, capaz de generar los árboles de dependencias, obtenida automáticamente a partir del corpus.
- **Base de datos** de colocaciones sintácticas de español, con información **estadística**, obtenida automáticamente a partir del corpus etiquetado sintacticamente, de manera supervisada.

1.6 Organización de la tesis

El resto del documento se organiza de la siguiente manera:

En el Capítulo 2 El problema de la ambigüedad sintáctica, se presenta una breve revisión del estado del arte sobre el lenguaje, procesamiento de lenguaje natural y ambigüedad. Se explica a detalle la ambigüedad sintáctica y cómo usar colocaciones para resolverla.

En el Capítulo 3 Los formalismos sintácticos de constituyentes y de dependencias, se describen algunos formalismos de los dos enfoques principales en el análisis sintáctico, así como los formalismos mixtos. Al final se hace una comparación de todos ellos.

En el Capítulo 4 Transformación del corpus de constituyentes en un corpus de dependencias, se describe el algoritmo utilizado para transformar el corpus orientado a constituyentes Cast3LB a un corpus de dependencias. Se presenta la implementación de tal algoritmo y la metodología que utilizamos para evaluarlo, así como los resultados obtenidos.

En el Capítulo 5 Extracción del diccionario de colocaciones del corpus, se presenta una breve descripción sobre colocaciones, tipos de colocaciones y sus aplicaciones. Además se describe el algoritmo utilizado para extraer el diccionario de colocaciones del corpus. Se presenta la implementación de tal algoritmo, así como los resultados obtenidos.

En el Capítulo 6 Evaluación del parser de dependencias DILUCT, se explica el funcionamiento del parser DILUCT y el algoritmo de evaluación de éste utilizando nuestro diccionario de colocaciones. Al final se muestran los resultados obtenidos.

En el Capítulo 7 Conclusiones y trabajo futuro, se presenta una discusión de lo visto en esta tesis, las conclusiones, aportaciones de la tesis, publicaciones generadas y trabajo futuro.

Finalmente se agrega un **glosario** con definiciones de términos, palabras, siglas, etc. que pueden contribuir al entendimiento del documento. Así mismo se incluyen referencias, con las bibliografías consultadas y referenciadas que aparecen enumeradas y ordenadas alfabéticamente. Finalmente un **índice de términos** con la lista de la ubicación dentro del documento de algunos términos y temáticas ordenadas alfabéticamente para facilitar su localización.

CAPÍTULO

2

La estructura sintáctica en el lenguaje natural y el problema de la ambigüedad sintáctica

Capítulo 2. La estructura sintáctica en el lenguaje natural y el problema de la ambigüedad sintáctica

2.1 Introducción

La ambigüedad sintáctica es un problema que se presenta en todos los lenguajes naturales. Podríamos decir que para los seres humanos la ambigüedad en el lenguaje pasa desapercibida, debido a que la resolvemos casi inconcientemente utilizando la realidad en que vivimos, el contexto y el conocimiento que poseemos sobre algunos temas. Pero para las computadoras no es así, por ello el procesamiento de lenguaje natural tiene por objetivo capturar esta información, dando la funcionalidad necesaria a las computadoras para que puedan analizar y procesar lenguaje natural, y así, intentar comprender como lo hacemos las personas.

A lo largo de este capítulo daremos una breve revisión del estado del arte sobre el lenguaje y el procesamiento de lenguaje natural. Después hablaremos del problema de la ambigüedad, y más a detalle, la ambigüedad sintáctica y los modelos de conocimiento que se utilizan para resolverla. Al final hablaremos de como utilizar colocaciones para resolver este tipo de ambigüedad.

2.2 Lenguaje

Un lenguaje se considera como un conjunto frases y oraciones, generalmente infinito y que se forma mediante combinaciones de palabras definidas en un diccionario previamente establecido. Estas combinaciones deben estructurarse correctamente, es

decir, respetar un conjunto de reglas sintácticas; además deben tener sentido en un contexto dado, a lo que se denomina semántica.

A lo largo de la historia el ser humano ha utilizado el lenguaje para transmitir sus conocimientos, sentimientos, emociones, sensaciones, comunicarse con el resto de los humanos, ya sea de manera oral, gráfica, escrita o por señas.

Cuando hablamos de lenguajes se pueden diferenciar dos clases muy bien definidas: los lenguajes naturales (español, ruso, inglés, etc.) y los lenguajes formales (lenguajes de programación, lógica matemática, etc.)

2.2.1 Lenguaje Natural

Podríamos definir el lenguaje natural como el medio principal para la comunicación humana.

Con respecto a nuestro mundo, el lenguaje nos permite designar las cosas reales y razonar acerca de ellas, así como también crear significados. El lenguaje natural fue desarrollado y organizado a partir de la experiencia humana.

Los lenguajes naturales tienen un gran poder expresivo y pueden ser utilizados para analizar y razonar situaciones complejas.

Una propiedad única de los lenguajes naturales es la polisemia, es decir, la posibilidad de que una palabra en una oración tenga diversos significados. El carácter poli semántico de un lenguaje tiende a incrementar la riqueza de su componente semántico, haciendo casi imposible su formalización.

Podemos resumir que los lenguajes naturales se distinguen por las siguientes propiedades:

- Han sido desarrollados por enriquecimiento progresivo antes de cualquier intento de formación de una teoría.
- La importancia de su carácter expresivo es debida fundamentalmente a la riqueza del componente semántico (polisemia).
- Existe dificultad o imposibilidad de una formalización completa.

2.2.2 Lenguaje Formal

Un lenguaje formal es un lenguaje artificial compuesto por símbolos y formulas, que tiene como objetivo fundamental formalizar la programación de computadoras o representar simbólicamente el conocimiento científico.

Las palabras y oraciones en un lenguaje formal están perfectamente definidas, una palabra mantiene el mismo significado prescindiendo del contexto o su uso.

Los lenguajes formales son exentos de cualquier componente semántico fuera de sus operadores y relaciones, y es gracias a esta ausencia de significado especial que los lenguajes formales pueden ser usados para modelar una teoría de la mecánica, de la ingeniería eléctrica, en la lingüística u otra naturaleza. Esto equivale a decir que durante la concepción de lenguajes formales toda la ambigüedad es eliminada.

En resumen, las características de los lenguajes formales son las siguientes:

- Se desarrollan de una teoría preestablecida.
- Tienen componente semántico mínimo.
- Posibilidad de incrementar el componente semántico de acuerdo con la teoría a formalizar.
- La sintaxis produce oraciones no ambiguas.
- Los números tienen un rol importante.
- Poseen una completa formalización y por esto, potencialmente posibilitan la construcción computacional.

2.3 Procesamiento del Lenguaje Natural

El estudio del lenguaje está relacionado con varias disciplinas. Una de ellas es la lingüística general, que estudia la estructura general y descubre las leyes universales de funcionalidad de los lenguajes naturales. Estas estructuras y leyes, aunadas a los métodos computacionales forman la lingüística computacional.

La lingüística computacional puede ser considerada como un sinónimo de procesamiento de lenguaje natural, ya que la tarea principal de ésta es la construcción de programas que procesen palabras y textos en lenguaje natural [\[Bolshakov & Gelbukh, 2004\]](#).

Para llevar a cabo esta tarea, los sistemas de procesamiento de lenguaje natural deben tener conocimiento acerca de la estructura del lenguaje, para así poder pasar de texto a significado y viceversa. Los niveles que componen esta estructura se explican en el siguiente punto.

2.4 Niveles del Lenguaje

La lingüística general comprende 5 niveles principales para análisis de estructura del lenguaje [\[Bolshakov & Gelbukh, 2004\]](#) que son:

- a) Nivel Fonológico: trata de los sonidos que componen el habla, permitiendo formar y distinguir palabras.
- b) Nivel Morfológico: trata sobre la estructura de las palabras y las leyes para formar nuevas palabras a partir de unidades de significado más pequeñas llamadas morfemas.
- c) Nivel Sintáctico: trata sobre cómo las palabras pueden unirse para construir oraciones y cuál es la función que cada palabra realiza en esa oración.
- d) Nivel Semántico: trata del significado de las palabras y de cómo se unen para dar significado a una oración.
- e) Nivel Pragmático: estudia la intención del hablante al producir oraciones específicas o textos en una situación específica.

2.5 Ambigüedad

La ambigüedad, en el proceso lingüístico, se presenta cuando pueden admitirse distintas interpretaciones a partir de una representación dada o cuando existe confusión al tener diversas estructuras y no tener los elementos necesarios para eliminar las eventualmente incorrectas. Para desambiguar, es decir, para seleccionar los significados o las estructuras más adecuadas de un conjunto conocido de posibilidades, se requieren diversas estrategias de solución según el caso [\[Galicia-Haro, 99\]](#).

Debido a que existe ambigüedad aún para los humanos, su solución no es sólo lograr la asignación del sentido único por palabra en el análisis de textos, sino eliminar la gran cantidad de variantes que normalmente existen. La ambigüedad es el problema más importante en el procesamiento de textos en lenguaje natural, por lo que su resolución es la tarea más importante a llevar a cabo.

2.5.1 Tipos de ambigüedad

Se distinguen tres tipos principales de ambigüedad: léxica, semántica y sintáctica o estructural.

La ambigüedad léxica se presenta cuando las palabras pueden pertenecer a diferentes categorías gramaticales, por ejemplo *bajo* puede ser una preposición, un sustantivo, un adjetivo o una conjugación del verbo *bajar*.

La ambigüedad semántica se presenta cuando las palabras tienen múltiples significados, por ejemplo la palabra *banco* puede significar banco de peces, banco para tomar asiento o institución financiera.

La ambigüedad sintáctica, también conocida como ambigüedad estructural se presenta cuando una oración puede tener más de una estructura sintáctica. Por ejemplo de la oración “María habló con el profesor del instituto” se puede entender dos cosas diferentes: a) el profesor pertenece al instituto, o bien, b) el tema del que habló María con el profesor fue el instituto.

A continuación se explica más detalladamente la ambigüedad estructural y cómo resolverla.

2.6 Ambigüedad Sintáctica

Este tipo de ambigüedad ocurre cuando la información sintáctica no es suficiente para hacer una decisión de asignación de estructura, o como lo mencionamos anteriormente, cuando una oración puede tener mas de una estructura (árbol) sintáctica.

La ambigüedad existe aún para los hablantes nativos, es decir, hay diferentes lecturas para una misma frase. Por ejemplo, en la oración “*Juan vio a un hombre con un telescopio*”, puede pensarse que: Juan utilizó el telescopio para ver al hombre, o, el hombre visto por Juan tiene un telescopio. Los árboles sintácticos extraídos de esta oración se muestran en la figura 1.

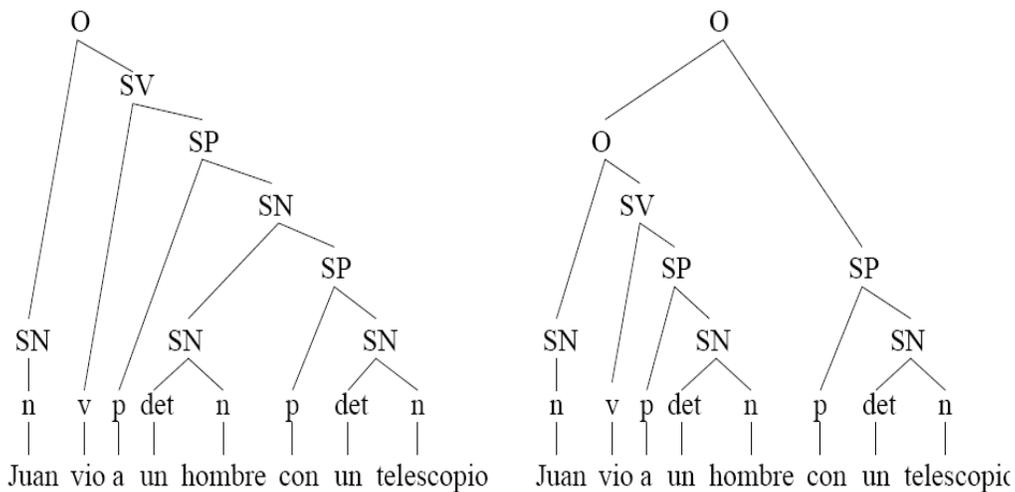


Figura 1. Árboles sintácticos extraídos de la oración “Juan vio a un hombre con un telescopio”.

La resolución de la ambigüedad sintáctica debe basarse en modelos de conocimiento diverso [Galicia-Haro, 99]. Tres de esos modelos se describen a continuación.

Patrones de manejo

Este método se basa en conocimiento lingüístico que adquieren los hablantes nativos durante el aprendizaje de su lenguaje. Este método es el más práctico para solucionar la mayoría de los problemas de ambigüedad. Aunque por sí mismo, este método no es suficiente para el análisis sintáctico de textos sin restricciones.

El conocimiento descrito en este modelo es la información léxica de verbos, adjetivos y algunos sustantivos del español, para enlazar las frases que realizan las valencias. No es posible establecer ese conocimiento mediante reglas o algoritmos pero es posible obtener la información léxica a partir de un corpus.

Reglas ponderadas

Es uno de los modelos de resolución de ambigüedad sintáctica más simple pero mucho más cómodo para aplicar y para compilar los recursos necesarios. Se trata de la utilización del formalismo de constituyentes (gramáticas generativas). Se codifica directamente el conocimiento gramatical en reglas de reescritura, es decir en gramáticas libres del contexto.

El conocimiento que se describe en este modelo es la clasificación y segmentación de la oración conforme a las categorías gramaticales de las palabras que la forman. La gramática está formada por un conjunto de reglas y por un conjunto de palabras, corresponde al lenguaje particular, ya que toda gramática es una teoría acerca de un lenguaje y por lo tanto no existen en ella descripciones neutrales.

Proximidad semántica

Este modelo está relacionado con el conocimiento semántico. Se requiere para desambiguar oraciones completas, porque sus diversas estructuras sintácticas son perfectamente posibles, o para enlazar frases circunstanciales que al no estar directamente enlazados con el sentido del lexema rector requieren un método conectado con la semántica de contexto.

Así que el conocimiento que describe es una clase de conocimiento semántico de contexto. Se trata de reconocer las palabras que están relacionadas, es decir, que están “más cercanas” semánticamente o que son “semánticamente compatibles”. Por ejemplo, en la frase conocida *Veo un gato con un telescopio* no es claro si *telescopio* está relacionado con *ver* o con *gato*. La información semántica permite decidir que *telescopio* está más próximo semánticamente de *ver* y no de la de *gato*.

La idea del empleo de la red semántica es la siguiente, por ejemplo, consideremos las frases: *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 qué 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*. De esta forma se desambiguan los enlaces.

2.7 Uso de colocaciones para resolver ambigüedad sintáctica

Este proceso de resolver ambigüedad sintáctica utilizando colocaciones se basa en el modelo de *patrones de manejo* –descrito anteriormente–, debido a que el conocimiento lingüístico que utiliza el sistema para elegir el árbol sintáctico correcto es el diccionario de colocaciones extraído de un corpus.

El algoritmo de 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 relaciones sintácticas en el diccionario de colocaciones, sumando las frecuencias de todas ellas. Si la colocación no se encuentra entonces su frecuencia es 0.
3. Se elige el árbol sintáctico que contenga la mayor suma de frecuencias de sus colocaciones.

Si el diccionario de colocaciones no tuviera frecuencias entonces se consideraría valor 0 si no existe la colocación y 1 si existe.

La idea presentada anteriormente para resolver ambigüedad sintáctica utilizando colocaciones no cubre los casos en que el número de relaciones sintácticas extraídas de los árboles sintácticos de una oración sea diferente.

2.8 Conclusiones

En este capítulo vimos que la ambigüedad se presenta en todos los niveles del lenguaje. Uno de los problemas principales de los sistemas de procesamiento de lenguaje natural es resolver la ambigüedad.

Los tres tipos principales de ambigüedad son léxica, semántica y sintáctica(o estructural). La ambigüedad sintáctica se presenta cuando se pueden obtener dos o más árboles sintácticos de una misma oración. Este tipo de ambigüedad es uno de los problemas más difíciles de resolver en sistemas de procesamiento de lenguaje natural.

Se presentó una idea de cómo resolver la ambigüedad sintáctica utilizando colocaciones, debido a que uno de las aportaciones de esta tesis es un diccionario de colocaciones.

CAPÍTULO

3

Los formalismos sintácticos de constituyentes y de dependencias

Capítulo 3. Los formalismos sintácticos de constituyentes y de dependencias

3.1 Introducción

Se tiende a creer que las palabras componen una oración como una progresión en una sola dimensión. Sin embargo, la propiedad del lenguaje natural, que es de importancia central en la sintaxis, es que tiene dos dimensiones. La primera es explícita, el orden lineal de palabras, y la segunda es implícita, la estructura jerárquica de palabras.

El orden lineal es lo mismo que la secuencia de las palabras en la oración. El papel de la estructura jerárquica se refiere a menudo como una dependencia. Podemos ejemplificarla con las siguientes frases:

Una persona sola en la construcción

Una persona interesada en la construcción

En la primera frase, el grupo de palabras *en la construcción* se une al grupo *una persona* indicando el lugar donde se encuentra la persona, mientras que en la segunda frase el mismo grupo se une a *interesada* indicando cuál es su interés. Lo que hace la diferencia en las interpretaciones, no es evidentemente un orden lineal puesto que el grupo *en la construcción* se encuentra en el final de ambas frases, y tampoco se trata de la distancia lineal en las dos frases.

Tanto el orden lineal como la estructura jerárquica, aunque principalmente esta última, son el tema principal en los formalismos para el análisis sintáctico que se presentan a continuación.

3.2 Formalismos de la lingüística computacional

Consideramos los dos enfoques que por mucho tiempo se han considerado opuestos: la gramática generativa (constituyentes) cuyo principal representante es la teoría desarrollada por Chomsky en sus diversas variantes, y la tradición estructuralista europea (dependencias) que proviene de Tesnière, con el ejemplo más representativo, la teoría Sentido \Leftrightarrow Texto de I. A. Mel'cuk.

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 [Chomsky, 65]. El trabajo que se realiza en la gramática generativa descansa en la suposición acerca de la estructura de la oración de 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 2 A en el árbol de constituyentes para la frase “*los niños pequeños estudian pocas horas*”.

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 O (oración), GN (grupo nominal), GV (grupo verbal), N (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.

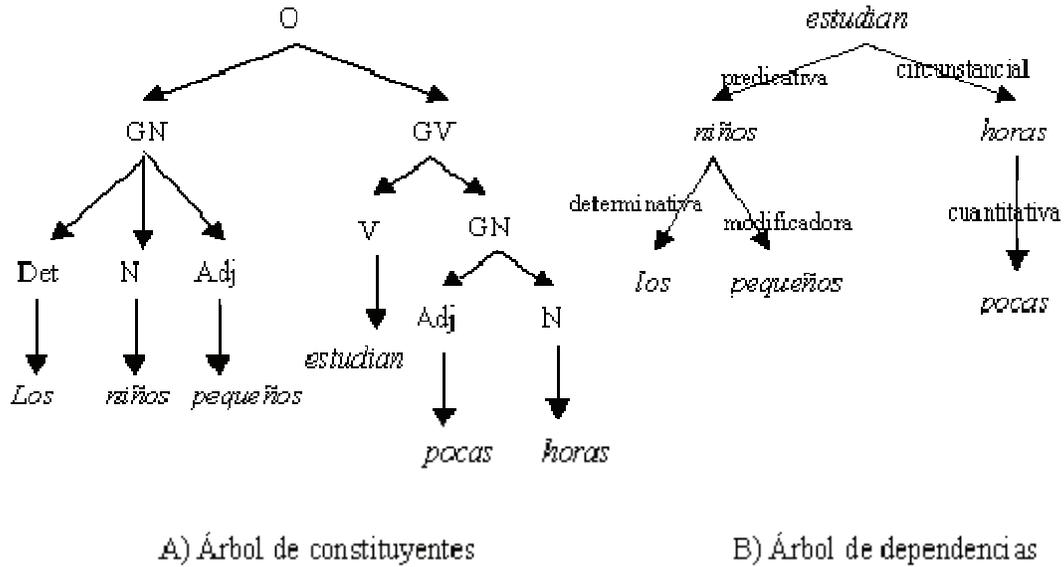


Figura 2. Estructuras sintácticas

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, el de [\[Tesnière, 59\]](#), 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 combinar 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 actantes) 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 *niñas*, la modificadora es *bonitas*. 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 2 B).

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 presentamos un panorama del desarrollo de éstos.

3.2.1 Formalismos de constituyentes

Chomsky [[Chomsky, 57](#)] presentó una versión inicial de la Gramática Generativa Transformacional (GGT), gramática en la cuál, 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 generar 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 2A:

O	→	GN GV	Adj	→	<i>pequeños pocas</i>
GV	→	V GN	Sust	→	<i>músicos horas</i>
GN	→	Art Sust Adj	V	→	<i>estudian</i>
GN	→	Adj Sust	Art	→	<i>los</i>

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 (estudian, los, 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 [Aho *et al.*, 86].

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 en base a los cambios propuestos en los trabajos de [\[Katz & Postal, 64\]](#) y de [\[Chomsky, 65\]](#). 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, y 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. [\[Salomaa, 71\]](#) y [\[Peters & Ritchie, 73\]](#) demostraron que el modelo transformacional era equivalente a una gramática sin restricciones.

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

En base a esto [\[Bresnan, 78\]](#) 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 transformacionales. Este paso de Bresnan fue seguido 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 [\[Montague, 70, 74\]](#), 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 éstos 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.

3.2.1.1 Government and Binding (GB)

Esta teoría apareció por primera vez en el libro *Lectures on Government and Binding* de [\[Chomsky, 82\]](#). El objetivo primordial de la GB, como mucho del trabajo de Chomsky, fue el desarrollo de una teoría de la gramática universal. La GB afirma que muchos de los principios que integran esta teoría están parametrizados, en el sentido de que los valores varían dentro de un rango limitado. La GB afirma que todos los lenguajes son esencialmente semejantes y que el conocimiento experimental con un lenguaje particular o con otro es una clase de fina sintonización dentro de un rango determinado, es decir, con unos pocos parámetros restringidos de posible variación.

La noción que adquiere un papel preponderante en el enfoque de constituyentes es una noción muy importante de la Gramatical Universal, la restricción. La suposición en que se basa esta teoría y que es compartida por muchas otras, es que cualquier cosa es posible y que los datos faltantes en la oración reflejan la operación de alguna restricción. El área más activa de investigación sintáctica desde los inicios de los ochenta ha sido precisamente resolver los detalles de este programa ambicioso.

En la GB se sigue el desarrollo del estilo modular de la EST, dividiendo la teoría de la gramática en un conjunto de subteorías, cada una con su propio conjunto universal de principios. Aunque la GB aún utiliza las derivaciones transformacionales para analizar oraciones, reduce la componente transformacional a una sola regla (Move- α), que puede mover cualquier elemento a cualquier lugar. La idea es que los principios generales filtren la mayoría de las derivaciones, previniendo la sobregeneración masiva que pudiera ocurrir.

3.2.1.2 Generalized Phrase Structure Grammar (GPSG)

La Gramática de Estructura de Frase Generalizada (Generalized Phrase Structure Grammar, GPSG) fue iniciada por Gerald Gazdar en 1981, y desarrollada por él y un grupo de investigadores, integrando ideas de otros formalismos; la teoría se expone detalladamente en [\[Gazdar *et al.*, 85\]](#).

La idea central de la GPSG es que las gramáticas usuales de estructura de frase independientes del contexto pueden mejorarse en formas que no enriquecen su capacidad generativa pero que las hacen adecuadas para la descripción de la sintaxis de lenguajes naturales. Al situar la estructura de frase, otra vez, en un lugar principal consideraban que los argumentos que se habían aducido contra las CFG, como una teoría de sintaxis, eran argumentos relacionados con la eficiencia o la elegancia de la notación y no realmente en cuanto a la cobertura del lenguaje.

La GPSG propone sólo un nivel sintáctico de representación que corresponde a la estructura superficial, y reglas que no son de estructura de frase en el sentido en que no están en una correspondencia directa con partes del árbol. Entre otras ideas importantes originadas en la teoría está la separación de las reglas en reglas de dominancia inmediata (reglas ID, Immediate dominance) que especifican solamente las frases que pueden aparecer como nodos en un árbol sintáctico, y las reglas de precedencia lineal (reglas LP, Linear precedence) que especifican restricciones generales que determinan el orden de los nodos en cualquier árbol.

Una consideración importante en las reglas, es que puede describirse información gramatical. Esta información gramatical codificada se toma como restricción en la admisibilidad en los nodos.

Esta teoría incluye la consideración del h-núcleo en las reglas, y de categorías. Las categorías son un conjunto de pares característica - valor. Las características tienen dos propiedades: tipos de valores y regularidades distribucionales (compartidos con otras características). La GPSG es de hecho una teoría de cómo la información sintáctica fluye dentro de la estructura. Esta información está codificada mediante características sintácticas. Todas las teorías sintácticas emplean características en diferentes grados, pero en la GPSG se emplean principios para el uso de características.

Los principios determinan cómo se distribuyen las características en el árbol, o restringen la clase de categorías posibles.

Otra idea importante en la GPSG es el tratamiento de las construcciones de dependencia a largas distancias, incluyendo las construcciones de llenado de faltantes (filling gap) como: topicalización, preguntas con Wh y cláusulas relativas. Este fenómeno estaba considerado como totalmente fuera del alcance de las gramáticas sin transformaciones. En las dependencias a larga distancia, sin límite, existe una relación entre dos posiciones en la estructura sintáctica, relación que puede alargarse.

El resultado más importante del análisis en la GPSG es que pudo manejar construcciones que se pensaba sólo podían describirse con la ayuda de las transformaciones. En este formalismo las transformaciones no figuran en ningún sentido en la teoría; es más, sin transformaciones de las dependencias de llenado de faltantes tuvo éxito en estos fenómenos donde la teoría transformacional había fallado.

3.2.1.3 Lexical-Functional Grammar (LFG)

La teoría de la Gramática Léxica Funcional (Lexical Functional Grammar, LFG) desarrollada por [\[Bresnan, 82\]](#) y [\[Dalrymple et al, 95\]](#) comparte con otros formalismos la idea de que conceptos relacionales, como sujeto y objeto, son de importancia central y no pueden definirse en términos de estructuras de árboles. La LFG considera que hay más en la sintaxis de lo que se puede expresar con árboles de estructura de frase, pero también considera la estructura de frase como una parte esencial de la descripción gramatical.

La teoría se ha centrado en el desarrollo de una teoría universal de cómo las estructuras de constituyentes se asocian con los objetos sintácticos. La LFG toma esos objetos sintácticos como primitivas de la teoría, en términos de las cuales se establecen una gran cantidad de reglas y condiciones.

En la LFG, hay dos niveles paralelos de representación sintáctica: la estructura de constituyentes (estructura-c) y la estructura funcional (estructura-f). La primera tiene la forma de árboles de constituyentes independientes del contexto. La segunda es un conjunto de pares de atributos y valores donde los atributos pueden ser características como tiempo y género, u objetos sintácticos como sujeto y objeto. En la LFG se

considera que la estructura-f despliega los objetos sintácticos. De esta manera se establecen las relaciones entre estructuras.

En la LFG cada frase se asocia con estructuras múltiples de distintos tipos, donde cada estructura expresa una clase diferente de información acerca de la frase. Siendo las dos representaciones principales las mencionadas estructura funcional y estructura de constituyentes (similar a la estructura superficial de la ST). Los principios generales y las restricciones de construcción específica definen las posibles parejas de estructuras funcionales y de constituyentes. La LFG reconoce un número más amplio de niveles de representación. Tal vez los más notables entre éstos son las estructuras- σ , que representan aspectos lingüísticamente relevantes del sentido, y la estructura-a que sirve para enlazar argumentos sintácticos con aspectos de sus sentidos [Bresnan, 95] y que codifica información léxica acerca del número de argumentos, su tipo sintáctico y su organización jerárquica, necesarios para realizar el mapeo a la estructura sintáctica.

Todos los elementos léxicos se insertan en estructuras-c en forma totalmente flexionada. Debido a que en la LFG no hay transformaciones, mucho del trabajo descriptivo que se hacía con transformaciones se maneja mediante un diccionario enriquecido, una idea importante de la LFG. Por ejemplo, la relación activa-pasiva se determina solamente por un proceso léxico que relaciona formas pasivas del verbo a formas activas, la cuál en lugar de tratarse como una transformación se maneja en el diccionario como una relación léxica entre dos formas de verbos.

En las LFG iniciales, la relación activa-pasiva fue codificada en términos de reglas léxicas, trabajo subsecuente ha buscado desarrollar una concepción más abstracta de las relaciones léxicas en términos de una teoría de mapeo léxico (TML). La TML provee restricciones en la relación entre estructuras-f y estructuras-a, es decir, restricciones asociadas con argumentos particulares que parcialmente determinan su función gramatical. Contiene también mecanismos con los cuales los argumentos pueden suprimirse en el curso de la derivación léxica. En la LFG la información de las entradas léxicas y las marcas de la frase se unifican para producir las estructuras funcionales de expresiones complejas.

3.2.1.4 Head-Driven Phrase Structure Grammar (HPSG)

La Gramática de Estructura de Frase dirigida por el h-núcleo (Head-driven Phrase Structure Grammar, HPSG) iniciada en [\[Pollard & Sag, 87\]](#) y revisada en [\[Pollard & Sag, 94\]](#) evolucionó directamente de la GPSG, para modificarla incorporando otras ideas y formalismos de los años ochenta. El nombre se modificó para reflejar el hecho de la importancia de la información codificada en los núcleos-h léxicos de las frases sintácticas, es decir, de la preponderancia del empleo de la marca *head* en el subconstituyente *hijo* principal.

En la HPSG se consideró que no había nada de especial en los sujetos salvo que era el menos oblicuo de los complementos que el h-núcleo selecciona. Para la GB el sujeto difiere de los complementos en la posición que tiene en el árbol de proyecciones. Esta consideración empezó a cambiar en la revisión de 1994 de la HPSG, basándose en los trabajos de [\[Borsley, 90\]](#), donde se considera el sujeto en forma separada.

La HPSG en [\[Pollard & Sag, 94\]](#) amplía el rango de los tipos lingüísticos considerados, los signos consisten no solamente de la forma fonética sino de otros atributos o características, con la finalidad de tratar una mayor cantidad de problemas empíricos. En esta teoría los atributos de la estructura lingüística están relacionados mediante una estructura compartida. De acuerdo a principios especiales introducidos en la teoría, las características principales de los h-núcleos y algunas de las características de los nodos hijos se heredan a través del constituyente abarcador.

El principal tipo de objeto en la HPSG es el signo (correspondiente a la estructura de características clase sign), y lo divide en dos subtipos disjuntos: los signos de frase (tipo frase) y los signos léxicos (tipo palabra). Las palabras poseen como mínimo dos atributos: uno fonético PHON (representación del contenido de sonido del signo) y otro SYNSEM (compuesto de información lingüística tanto sintáctica como semántica). Con los atributos y valores de estos objetos se crea una estructura de características.

De acuerdo a principios especiales introducidos en la teoría, las características principales de los h-núcleos y algunas de las características de los nodos hijos se heredan a través del constituyente abarcador.

Las frases tienen un atributo DAUGHTERS (DTRS), además de PHON y SYNSEM, cuyo valor es una estructura de características de tipo estructura de constituyentes (con-struct) que representa la estructura de constituyentes inmediatos de la frase. El tipo con-struct tiene varios subtipos caracterizados por las clases de hijos que aparecen en la frases. El tipo más simple y más empleado es el head-struct que incluye HEAD-DAUGHTERS (HEAD-DTR) y COMPLEMENT-DAUGHTERS (COMP-DTRS), que a su vez tienen atributos PHON y SYNSEM.

Un punto importante en la HPSG es que tiene varios principios: de constitución inmediata de las frases (proyección de los núcleos-h), de subcategorización, de semántica, etc., que realmente son restricciones disyuntivas. En la HPSG se considera que hay dos tipos de restricciones: de la gramática universal y de la gramática particular. Así que las expresiones gramaticales de un lenguaje particular dependen de las interacciones entre un sistema complejo de restricciones universales y particulares.

Para tratar los diversos fenómenos que en la GPSG se consideraron como dependencias sin límite, la HPSG emplea dos principios de la gramática universal (de realización de argumentos y el principio de faltantes) y una restricción del lenguaje particular (la condición sujeto).

En la HPSG, el diccionario, un sistema de entradas léxicas, corresponde a restricciones de la gramática particular. Cada palabra en el diccionario tiene información semántica que permite combinar el sentido de palabras diferentes en una estructura coherente unida.

Algunas de las ideas clave en la HPSG son entonces:

- 1) Arquitectura basada en signos lingüísticos.
- 2) Organización de la información lingüística mediante tipos, jerarquías de tipos y herencia de restricciones.
- 3) La proyección de frases mediante principios generales a partir de información con abundancia léxica.
- 4) Organización de esa información léxica mediante un sistema de tipos léxicos.

- 5) Factorización de propiedades de frases en construcciones específicas y restricciones más generales.

3.2.2 Formalismos de dependencias

[[Mel'cuk, 79](#)] 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 cuáles otros y en que 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) [[Perlmutter, 83](#)] 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 rearrreglo estructural de grupos nominales.

Muy pocas gramáticas de dependencia han sido desarrolladas recientemente [[Fraser, 94](#)], [[Lombardi & Lesmo, 98](#)]. A continuación, describimos los formalismos más representativos: Dependency Unification Grammar (DUG), Word Grammar (WG) y Meaning \leftrightarrow Text Theory (MTT).

3.2.2.1 Dependency Unification Grammar (DUG)

La historia de la Gramática de Unificación de Dependencias (Dependency Unification Grammar en inglés) [[Hellwig, 86](#)] comienza al inicio de los años setenta con el desarrollo del sistema llamado PLAIN [[Hellwig, 80](#)] aplicando diferentes métodos para la sintaxis y la semántica, y combinando una descripción sintáctica basada en dependencias llamada Gramática de Valencia con Transformaciones para simular relaciones lógico semánticas. Desde los inicios empleó categorías complejas con atributos y valores, y un mecanismo de subsumisión para establecer la concordancia. En los años ochenta enfatizó su filiación a las gramáticas de unificación resultando en la DUG. Desde entonces tanto PLAIN como DUG se han aplicado en diversos proyectos y se han ido modificando.

La noción de unificación corresponde a la idea de unión de conjuntos, para la mayoría de los propósitos. La unificación es una operación para combinar o mezclar dos elementos en uno solo que concuerde con ambos. Esta operación tiene gran importancia en estructuras de rasgos (género, etc.). La unificación difiere en que falla si algún atributo está especificado con valores en conflicto, por ejemplo: al unificar dos atributos de número dónde uno es plural y otro es singular [[Briscoe & Carroll, 93](#)].

Tres conceptos son los más importantes en esta teoría como gramática de dependencias: el lexicalismo, los complementos y las funciones.

Por lexicalismo se considera la suposición de que la mayoría de los fenómenos en un lenguaje depende de los elementos léxicos individuales, suposición que es válida para la sintaxis (igualando los elementos léxicos con las palabras). Los complementos son importantes para establecer todas las clases de propiedades y relaciones entre objetos en el mundo verdadero. La importancia de las funciones entre

otras categorías sintácticas está relacionada con que cada complemento tiene una función específica en la relación semántica establecida por el h-núcleo. La función concreta de cada complemento establece su identidad y se hace explícita por una explicación léxica, por ejemplo: el verbo persuadir requiere un complemento que denote al persuasor, otro complemento que denote la persona persuadida y aún otro que denote el contenido de la persuasión.

En la DUG, una construcción sintáctica estándar consiste de un elemento h-núcleo y un número de constituyentes que completan a ese elemento h-núcleo. Para este propósito se necesitan palabras que denoten la propiedad o relación, y expresiones que denoten las entidades cualificadas o relacionadas. La morfología y el orden de palabras marcan los roles de los constituyentes respectivos en una oración. En ausencia de complementos, el rector, es decir el verbo, está insaturado. Sin embargo, es posible predecir el número y la clase de construcciones sintácticas que son adecuadas para complementar cada palabra rectora particular.

Como la DUG se ha aplicado principalmente al alemán considera el orden de palabras en el árbol de dependencias. Este árbol difiere de los árboles usuales de gramáticas de dependencias en que los nodos tienen etiquetas múltiples. El orden de palabras es entonces otro atributo. Se examina el orden lineal de los segmentos que se asocian a los nodos del árbol de dependencias. DUG considera características de posición con valores concretos que se calculan y se sujetan a la unificación.

3.2.2.2 Word Grammar (WG)

La Gramática de Palabra (Word Grammar, WG), para su autor [\[Hudson, 84\]](#), es una teoría general de la estructura del lenguaje y aunque sus bases son lingüísticas y más específicamente gramaticales, considera que su mayor intención es contribuir a la psicología cognitiva ya que ha desarrollado la teoría desde el inicio con el propósito de integrar todos los aspectos del lenguaje en una teoría que sea compatible con lo que se conoce acerca de la cognición general, aunque este objetivo no se ha logrado todavía.

Hudson ve la WG como una teoría del lenguaje en forma cognitiva, como una red que contiene tanto la gramática como el diccionario y que integra el lenguaje con el resto de

la cognición. La semántica en WG sigue a [\[Lyons, 77\]](#), [\[Halliday, 67, 68\]](#) y [\[Fillmore, 76\]](#) en lugar de seguir la lógica formal.

La suposición de la WG es que el lenguaje puede analizarse y explicarse en la misma forma que otras clases de conocimiento o comportamiento. Como su nombre lo sugiere, la unidad central de análisis es la palabra. Las palabras son las únicas unidades de la sintaxis, y la estructura de la oración consiste totalmente de las dependencias entre palabras individuales. Por lo que la WG es claramente parte de la tradición de gramáticas de dependencias.

Una segunda versión, la English Word Grammar (EWG) [\[Hudson, 90\]](#) introduce cambios importantes para detallar el análisis, la terminología y la notación, en lo que concierne a la teoría sintáctica, con la adición de estructura superficial y la virtual abolición de características.

La mayor parte del trabajo en la WG trata de la sintaxis aunque también se ha desarrollado cierto trabajo en la semántica y algo más tentativo en la morfología [\[Hudson, 98\]](#). Para la WG las palabras no nada más son las unidades más grandes de la sintaxis sino que también son las unidades más pequeñas por lo que las estructuras sintácticas no pueden separar bases e inflexiones, esto hace que la WG sea un ejemplo de sintaxis independiente de la morfología.

3.2.2.3 Meaning ⇔ Text Theory (MMT)

La Teoría Texto ⇔ Significado (Meaning ⇔ Text Theory, MTT), desde el ensayo en la publicación [\[Mel'cuk & Zholkovsky, 70\]](#) ha sido elaborada y refinada en diversos artículos y libros. La concepción de cómo los significados léxicos interactúan con las reglas sintácticas es de las mejor desarrolladas y con más principio en la literatura.

La meta de la teoría es modelar la comprensión del lenguaje como un mecanismo que convierta los significados en los textos correspondientes y los textos en los significados correspondientes. Aunque no hay una correspondencia de uno a uno, ya que el mismo significado puede expresarse mediante diferentes textos, y un mismo texto puede tener diferentes significados.

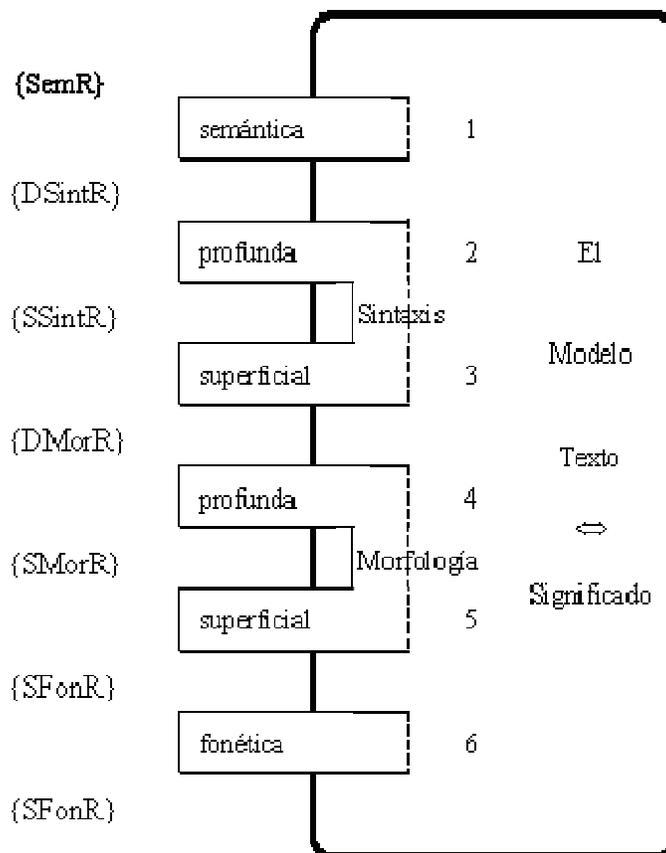


Figura 3. Niveles de Representación en la MTT.

La MTT emplea un mayor número de niveles de representación, tanto la sintaxis como la morfología y la fonología se dividen en dos niveles: profundo (D) y superficial (S). Bajo estos términos, la morfología profunda (DMorR) es más superficial que la sintaxis superficial (SSintR). Las nociones de profundo y más superficial significan que conforme progresa la representación de la semántica a la fonología superficial (SFonR) se vuelve más y más, detallada y específica del lenguaje.

La MTT es un sistema estratificado. Cada oración se caracteriza simultáneamente por siete diferentes representaciones, cada una especifica la oración desde la perspectiva del nivel correspondiente. Cada nivel de representación se mapea al adyacente mediante una de las seis componentes de la MTT. En la figura 3 se muestran estos siete niveles como en [\[Mel'cuk, 88\]](#).

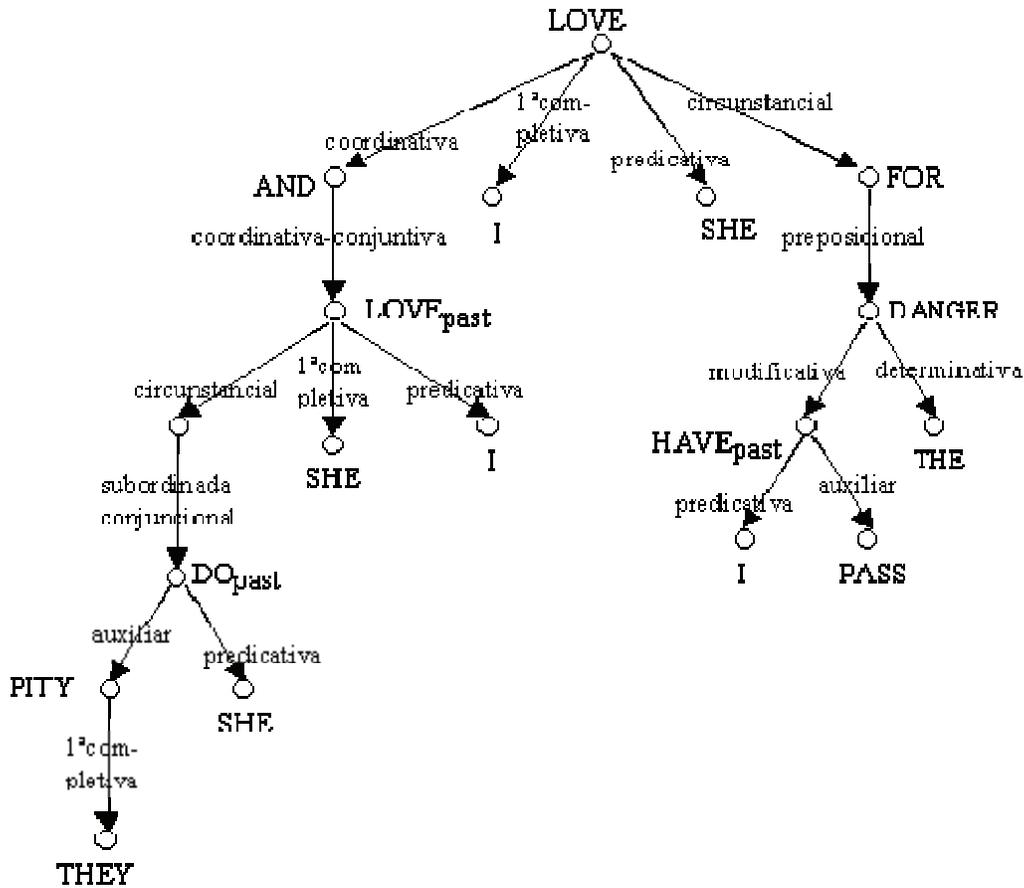


Figura 4. Ejemplo de estructura de dependencias en la MTT

En la Figura 4, se presenta un ejemplo del árbol de dependencias de acuerdo a la MTT de [Mel'cuk, 88] para la frase “*She lov'd me for the dangers I had pass'd, and I lov'd her that she did pity them*” (en español, *Ella me ama por los peligros que yo he pasado y yo la amo por la lástima que ella les tiene*), donde se hace una comparación con un árbol de constituyentes. Este árbol de dependencias presenta dos ventajas: requiere exactamente dieciocho nodos (el número de palabras), el orden lineal de los nodos es absolutamente irrelevante ya que la información se preserva a través de las dependencias etiquetadas.

Cada nivel de representación se considera como un lenguaje separado en el sentido de que tiene su propio vocabulario diferente y reglas distintas de combinación. La transición de un nivel a otro es un proceso de tipo traducción que involucra el cambio tanto de los elementos como de las relaciones entre ellos, pero que no cambia el contenido informativo de la representación.

Tres conjuntos de conceptos y términos son esenciales en la MTT en su aproximación a la sintaxis:

- Una situación y sus participantes (actuantes).
- Una palabra y sus actuantes semánticos que forman la valencia semántica de la palabra.
- Una palabra y sus actuantes sintácticos que forman la valencia sintáctica de la palabra.

La situación, en esta teoría, significa un bloque de la realidad reflejada por la lexis de un lenguaje dado. Los actuantes semánticos de una situación deben y pueden determinarse sin ningún recurso de la sintaxis y corresponden a esas entidades cuya existencia está implicada por su significado léxico. Por ejemplo, para [\[Mel'cuk, 88\]](#) la diátesis es la correspondencia entre los actuantes: semánticos, de la sintaxis profunda, y de la sintaxis superficial.

Los actuantes semánticos y los roles temáticos son similares aunque los roles temáticos siguiendo la tradición de los constituyentes tratan de generalizar los participantes y la MTT los particulariza, describiéndolos para cada verbo específico.

La MTT usa la noción de valencia sintáctica, es decir, la totalidad de los actuantes sintácticos de la palabra, esta noción es similar a la característica de subcategorización de la vieja gramática transformacional y a los argumentos de la teoría X-barra. La diferencia es que la valencia sintáctica se define independientemente de, y en yuxtaposición a, la valencia semántica. Esto hace posible usar claramente consideraciones semánticamente especificadas en la definición de la valencia sintáctica y marcar una diferencia entre ellas y las consideraciones sintácticas.

3.2.3 Formalismos mixtos

En los años setenta los términos lexicismo y lexicalismo se utilizaron para describir la idea de emplear reglas léxicas para capturar fenómenos que eran analizados previamente por medio de transformaciones. Por ejemplo, mediante una regla léxica se podía obtener a partir de un verbo una forma de adjetivo, de pelear obtener peleonero. Por lo que se establecía que las reglas sintácticas no debían hacer referencia a la

composición interna morfológica. El lexicalismo ahora, en forma muy burda, puede considerarse como una aproximación para describir el lenguaje, que enfatiza el diccionario a expensas de las reglas gramaticales.

Resulta engañosa esta caracterización inicial porque el lexicalismo cubre un rango amplio de aproximaciones y teorías que capturan este énfasis léxico en formas muy diferentes. Por ejemplo, dos enfoques principales son: que tanta información como sea posible acerca de la buena formación sintáctica esté establecida en el diccionario, y que las reglas sintácticas no deben manipular la estructura interna de las palabras.

El lexicalismo estricto para [\[Sag & Wasow, 99\]](#) es que las palabras, formadas de acuerdo a una teoría léxica independiente, son los átomos de la sintaxis. Su estructura interna es invisible a las restricciones sintácticas. Para él, el lexicalismo radical define que todas las reglas gramaticales se ven como generalizaciones sobre el diccionario. El principio de lexicalismo estricto, para este autor, tiene su origen en el trabajo de [\[Chomsky, 70\]](#), quien desafió los intentos previos para derivar nominalizaciones (por ejemplo, la compra de una pelota por el niño) a partir de cláusulas (por ejemplo, el niño compró una pelota) vía transformaciones sintácticas.

Aunque el lexicalismo originalmente se vio relacionado con la reducción de potencia y capacidad de las reglas transformacionales, actualmente se ve de una forma más general relacionada a la reducción de la potencia y capacidad de las reglas sintácticas de cualquier clase, y por lo tanto con un énfasis mayor en los diccionarios.

Los formalismos de constituyentes en su evolución han ido modificando conceptos que los aproximan a las dependencias. La LFG mantuvo la representación de estructura de frase para representar la estructura sintáctica de superficie de una oración, pero tuvo que introducir la estructura funcional para explicar explícitamente los objetos sintácticos, la cuál es esencialmente una especificación de relaciones de dependencia sobre el conjunto de lexemas de la oración que se describe.

La RG constituye una desviación decisiva de la estructura de frase hacia las dependencias, al establecer que los objetos sintácticos deben considerarse como nociones primitivas y deben figurar en las representaciones sintácticas. La relación

gramatical como ser el sujeto de, o ser el objeto directo de es una clase de dependencia sintáctica.

La HPSG, en su última versión [[Sag & Wasow, 99](#)] está formulada en términos de restricciones independientes del orden. Como heredera del enfoque de constituyentes incluye restricciones en sustitución de las transformaciones, pero se basa en la observación de la reciente literatura psicolingüística de que el procesamiento lingüístico humano de la oración tiene una base léxica poderosa: las palabras tienen una información enorme, por lo que ciertas palabras clave tienen un papel de pivotes en el procesamiento de las oraciones que las contienen, esta noción está presente en la MTT desde sus inicios. También la Word Grammar [[Hudson 84](#)] y el Word Expert Parser [[Small, 87](#)] proclaman esta base psicolingüística.

Esta observación, modifica el concepto de estructura de frase en la HPSG, donde la noción de estructura de frase se construye alrededor del concepto h-núcleo léxico, una sola palabra cuya entrada en el diccionario especifica información que determina propiedades gramaticales cruciales de la frase que proyecta. Entre esas propiedades se incluye la información de POS (los sustantivos proyectan grupos nominales, los verbos proyectan oraciones, etc.) y relaciones de dependencias (todos los verbos requieren sujeto en el inglés, pero los verbos difieren sistemáticamente en la forma en que seleccionan complementos de objeto directo, complementos de cláusula, etc.).

El lexicalismo, a nuestro entender, representa la convergencia en los enfoques de constituyentes y de dependencias. Aunque las dependencias, desde su origen le han dado una importancia primordial a las palabras y a las relaciones léxicas entre ellas, el enfoque de constituyentes vía el lexicalismo considera, en sus versiones más recientes (por ejemplo la última revisión a la HPSG), muchos de los conceptos de aquellas.

3.3 Comparación de los formalismos sintácticos

A pesar de la discusión de cuarenta años en la literatura, no hay consenso en cuanto a cual formalismo es mejor. Aunque los formalismos combinados tales como HPSG [[Sag](#)

[et al., 2003](#)] que se han propuesto, parecen llevar la herencia de las ventajas así como las desventajas de ambos enfoques, éstas últimas impiden su uso amplio en la práctica del procesamiento de lenguaje natural. La utilización de uno de los dos acercamientos depende probablemente de la tarea a realizar.

Desde el punto de vista de implementación, los formalismos gramaticales tienen una importante influencia sobre la forma de representación de las frases, representaciones que son la base de todo el razonamiento posterior en los programas informáticos. Las gramáticas generativas son inadecuadas relativamente para este fin y no tuvieron aplicación real en informática. De entre ellas, la GPSG es la extensión más interesante por su ambición de tratar los aspectos semánticos.

En la evolución de las gramáticas generativas, éstas se tuvieron que aumentar para incluir la concordancia y en algunas versiones se consideró la unificación de los rasgos. Una característica fundamental de las gramáticas funcionales, como la LFG es que permiten integrar aspectos semánticos, en este sentido constituyeron uno de los ejes de investigación más importantes. Pusieron de relieve también la importancia primordial del léxico dentro de las descripciones lingüísticas.

Ninguno de los formalismos hasta ahora desarrollados abarca todos los fenómenos lingüísticos, es decir, no tiene una cobertura amplia del lenguaje. El fenómeno de dependencias lejanas motivó una cantidad significativa de investigación en los formalismos gramaticales. En la gramática generativa en su primera etapa, se manejaron fuera de la CFG. La LFG y la GPSG propusieron métodos de capturar las dependencias con el formalismo de CFG, empleando rasgos o características. Otra línea ha sido tratar de definir nuevos formalismos que sean más poderosos que la CFG y que puedan manejar dependencias lejanas, como las TAG.

La última tendencia es en formalismos más orientados hacia los mecanismos computacionales, como la HPSG, la CG, la DUG. Las dos primeras emplean información de subcategorización extensivamente y haciéndolo simplifican de manera significativa la CFG a expensas de un diccionario más complicado. En la DUG, como en las gramáticas de dependencias, se definen todos los objetos de las palabras por lo que los diccionarios son el elemento central ya que no se emplean reglas.

3.4 Conclusiones

Como vimos anteriormente, existen dos formalismos principales para la representación de la estructura sintáctica de una oración: constituyentes (o estructura de frase) y dependencias. Ambos tipos utilizan árboles para representar la estructura de frase de una oración, aunque el significado de los nodos y sus relaciones en el árbol son diferentes.

La representación de dependencias simplifica grandemente algunas tareas con respecto a la de constituyentes. Por ejemplo:

- En lexicografía, recopilar estadísticas sobre la combinación sintáctica de palabras individuales (*leer un libro y martillar un clavo* contra **leer un clavo y *martillar un libro*) es trivial bajo la representación de dependencias: sólo cuentas las frecuencias de los arcos que conectan los casos de dos palabras dadas en un corpus. Uno de muchos usos de tales estadísticas [[Bolshakov, 2004](#); [Bolshakov y Gelbukh, 2001, 2003](#)] es la desambiguación sintáctica: el árbol con combinaciones de pares de palabras más frecuentes es elegido [[Yuret, 98](#); [Gelbukh, 99](#)]. Con la representación de constituyentes esto es difícil o casi imposible.

- En recuperación de información y minería del texto, encontrar una frase o hacer una búsqueda compleja en un corpus es relativamente sencillo en un árbol de dependencias: la búsqueda *camisa con mangas largas y franjas rojas* se encontrará fácilmente en una descripción *Una camisa de seda de alta calidad con franjas verticales anchas rojas y mangas largas azules* en un base de datos de comercio electrónico, pero no con la descripción *Una camisa roja con franjas largas azules en las mangas*.

- En el análisis semántico, transformando el árbol de dependencias en cualquier representación semántica más cercana, como un grafo conceptual [[Sowa, 84](#)] o una red semántica [[Mel'čuk, 88](#)], es mucho más directo. De hecho, HPSG construye una clase de árbol de dependencias para hacer su representación semántica de repetición mínima [[Sag et al, 2003](#)].

CAPÍTULO

4

4

Transformación del corpus de constituyentes a un corpus de dependencias

Capítulo 4. Transformación del corpus de constituyentes a un corpus de dependencias

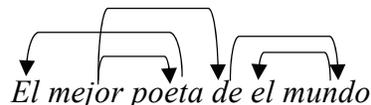
4.1 Introducción

En el capítulo anterior hablamos de los dos principales formalismos de representación sintáctica: constituyentes y dependencias.

La mayoría de herramientas y recursos existentes se orientan a la representación de constituyentes. La extracción de colocaciones (relaciones sintácticas) se hace en base a estructuras orientadas a dependencias, es por ello que hicimos un algoritmo para convertir el corpus (de constituyentes) en español Cast3LB en un corpus de dependencias.

A pesar de sus diferencias, ambas representaciones comparten el volumen de información en la estructura sintáctica -a tal grado que pueden ser combinadas [Sag et al, 2003] una puede ser derivada automáticamente de la otra, dando una cierta información que se agrega y está presente en la segunda representación pero ausente en la primera. Básicamente, la representación de constituyentes lleva más información sobre el orden de las palabras dentro de una unidad estructural, mientras que la representación de dependencias lleva más información sobre la herencia o relación de características sintácticas dentro de tal unidad.

De esta forma, la información que contiene un árbol de constituyentes se puede agregar automáticamente a un árbol de dependencias. Obviamente, tal conversión no puede ser totalmente exacta porque las dos representaciones están en desigualdad cuando existen construcciones gramaticales más raras tales como construcciones no-proyectivas, por ejemplo:



En el algoritmo que utilizamos para la transformación del corpus de constituyentes a un corpus de dependencias, y que se explica en este capítulo, no tomamos en cuenta tales detalles.

4.2 El corpus en español Cast3LB

Cast3LB es un corpus de cien mil palabras (aproximadamente 3,500 oraciones) creado a partir de dos corporas: 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 [[Navarro et al., 2003](#)].

4.3 Extracción de la gramática de constituyentes

Para extraer la gramática del corpus Cast3LB llevamos a cabo los siguientes pasos:

a) Simplificación del corpus de constituyentes.

Todas las etiquetas utilizadas por el corpus se componen de dos partes. La primera especifica la categoría gramatical, por ejemplo, sustantivo, verbo, grupo nominal, grupo verbal, oración, etc. Esta parte de la etiqueta es la más importante para la extracción de la gramática.

(oración
(S		frase nominal del sujeto
(sn-SUJ		especificador femenino plural
(espec.fp		determinante artículo femenino plural
(da0fp0 Las el))		grupo nominal femenino plural
(grup.nom.fp		sustantivo común femenino plural
(ncfp000 reservas reserva)		frase preposicional
(sp		preposición
(prep		adposición preposición simple
(sps00 de de))		frase nominal
(sn		grupo nominal coordinante
(grup.nom.co		grupo nominal masculino singular
(grup.nom.ms		sustantivo común masculino singular
(ncms000 oro oro))		coordinante
(coord		conjunción coordinada
(cc y y))		grupo nominal femenino plural
(grup.nom.fp		sustantivo común femenino plural
(ncfp000 divisas divisa))))))		frase preposicional
(sp		preposición
(prep		adposición preposición simple
(sps00 de de))		frase nominal
(sn		grupo nominal
(grup.nom		sustantivo propio
(np000000 Rusia Rusia))))))		grupo verbal
(gv		verbo principal indicat. pasado tercera pers. plural
(vmis3p0 subieron subir))		frase nominal
(sn-CC		grupo nominal
(grup.nom		cantidad monetaria
(Zm 800 millones de dolares 800 millones de dolares))))))		

Figura 5. Oración con etiquetas del corpus Cast3LB (“Las reservas de oro y divisas de Rusia subieron 800_millones_de_dolares”).

La segunda parte especifica datos adicionales como pueden ser género y número para frases nominales, o el tipo de la oración subordinada. Para simplificar el corpus y reducir el número de reglas gramaticales eliminamos la segunda parte de la etiqueta (datos adicionales), ya que al hacerlo no se afecta la estructura del corpus, por lo tanto la transformación tampoco. Por ejemplo, para una oración el corpus utiliza: *S*(oración), *S.F.C.*(oración coordinada) o *S.F.C.co-CD* (oración coordinada del objeto). Nosotros mapeamos todas ellas a una sola etiqueta, *S*. Para grupos nominales, Cast3LB utiliza *grup.nom* (grupo nominal), *grup.nom.fp* (grupo nominal femenino plural), *grup.nom.ms* (grupo nominal masculino singular), *grup.nom.co* (grupo nominal coordinado), etc.; nosotros mapeamos todas a la etiqueta *grupnom*. La figura 5 muestra una oración del corpus Cast3LB usando las etiquetas originales. La figura 6 muestra la misma oración ya

(oración
(S		frase nominal
(sn		especificador
(espec		determinante artículo
(da Las el))		grupo nominal
(grupnom		sustantivo
(n reservas reserva)		frase preposicional
(sp		preposición
(prep		adposición preposición simple
(sps00 de de))		frase nominal
(sn		grupo nominal
(grupnom		grupo nominal
(grupnom		sustantivo
(n oro oro))		coordinante
(coord		conjunción coordinada
(cc y y))		grupo nominal
(grupnom		sustantivo
(n divisas divisa))))))		frase preposicional
(sp		preposición
(prep		adposición preposición simple
(sps00 de de))		frase nominal
(sn		grupo nominal
(grupnom		sustantivo
(n Rusia Rusia))))))		grupo verbal
(gv		verbo principal
(vm subieron subir))		frase nominal
(sn		grupo nominal
(grupnom		cantidad monetaria
(Zm 800 millones de dolares 800 millones de dolares))))))		

Figura 6. Oración simplificada del corpus Cast3LB (“Las reservas de oro y divisas de Rusia subieron 800_millones_de_dolares”).

simplificada. Además, para reducir el número de patrones de la gramática resultante, simplificamos el corpus eliminando todos los signos de puntuación (puntos, comas, etc.).

b) Extracción de patrones.

Para extraer todas las reglas de la gramática, se recorre cada árbol del corpus en profundidad de izquierda a derecha, los nodos con más de un hijo se consideran como la parte izquierda de una regla, y sus hijos se consideran la parte derecha de la regla. Por ejemplo, los patrones extraídos de la figura 7 se muestran en la figura 8.

Los patrones extraídos son para nosotros las reglas gramaticales del corpus Cast3LB. Ejemplo, la regla *grupnom* ← *n sp sp* nos dice que un grupo nominal puede componerse de un sustantivo y una frase preposicional y otra frase preposicional y a su vez la regla *sp* ← *prep sn* nos dice que una frase preposicional se puede componer de una preposición y una frase nominal, etc.

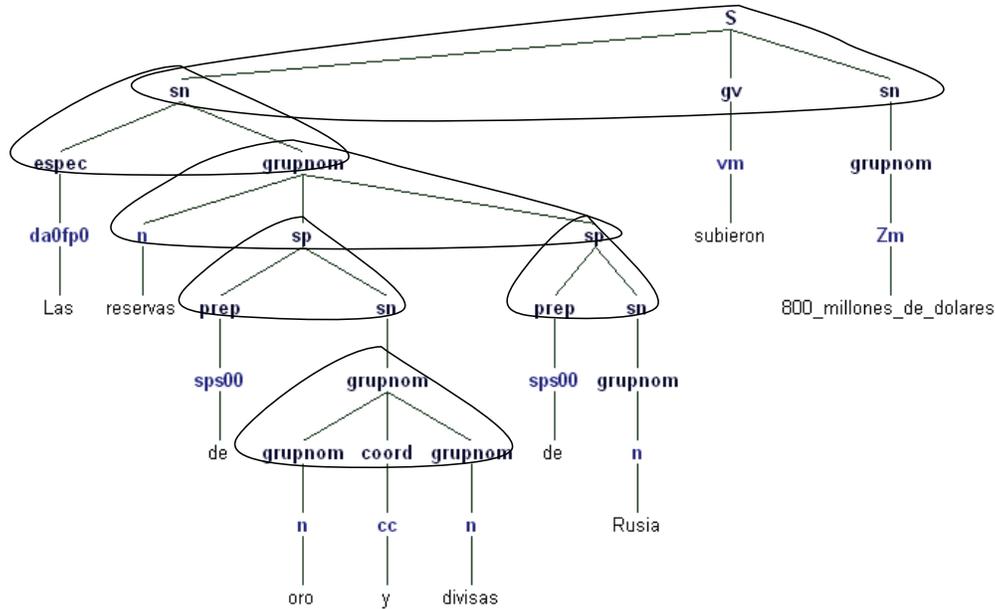


Figura 7. Patrones a extraer de la oración “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares”.

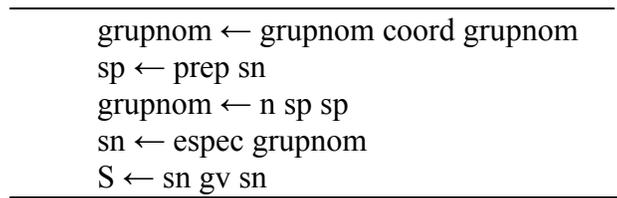


Figura 8. Patrones extraídos de la oración “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares”

4.4 Heurísticas para la determinación de los jefes de frases en las reglas de gramática

Para marcar automáticamente los jefes de frases (cabezas) en las reglas gramaticales, usamos reglas de heurística. Utilizamos el símbolo @ para denotar la cabeza de una regla (patrón). La heurística utilizada es la siguiente:

1. Si el patrón contiene un solo elemento como parte derecha entonces se marca como cabeza, ejemplo:

$$\text{grupnom} \leftarrow @ \text{ n}$$

Esta regla se lleva a cabo al momento de extraer la gramática del corpus de constituyentes, marcando como cabezas (en el mismo corpus) los nodos que solo tienen una hoja. Las siguientes reglas se aplican después de la extracción de la gramática.

2. Si el patrón contiene un coordinante (*coord*) entonces se marca como cabeza.

$$\text{grupnom} \leftarrow \text{grupnom @coord grupnom}$$
$$S \leftarrow @coord \text{sn gv sn}$$

3. Si el patrón contiene dos o más coordinantes (*coord*) entonces el primero es la cabeza,

$$S \leftarrow @coord S \text{ coord } S$$
$$Sp \leftarrow @coord \text{sp coord sp}$$

4. Si el patrón contiene una frase verbal (*gv*) entonces ésta es la cabeza,

$$S \leftarrow \text{sn @gv sn}$$
$$S \leftarrow \text{sadv sn @gv } S$$

5. Si el patrón contiene un pronombre relativo (*relatiu*), entonces éste se marca como cabeza,

$$\text{sp} \leftarrow \text{prep @relatiu}$$
$$\text{sn} \leftarrow @relatiu \text{ grupnom}$$

6. Si el patrón contiene una preposición (*prep*) entonces es marcada como cabeza, ejemplo:

$$\text{sp} \leftarrow \text{sadv @prep sn}$$
$$\text{sp} \leftarrow @prep \text{ sp}$$

7. Si el patrón contiene un verbo infinitivo (*infinitiu*) entonces es la cabeza,

$$S \leftarrow @infinitiu S \text{ sn}$$
$$S \leftarrow \text{conj @infinitiu}$$
$$S \leftarrow \text{neg @infinitiu sa}$$

8. Si el patrón contiene un participio presente (*gerundi*), entonces se marca como cabeza,

$$S \leftarrow @gerundi S$$

9. Si el patrón contiene un verbo principal (*vm*) entonces éste es la cabeza,

$$\text{gv} \leftarrow \text{va @vm}$$
$$\text{infinitiu} \leftarrow \text{va @vm}$$

10. Si el patrón contiene un verbo auxiliar (*va*) y cualquier otro verbo, entonces el verbo auxiliar nunca es la cabeza, ejemplo:

$$gv \leftarrow va @vs$$

11. Si el patrón contiene un especificador (*espec*) y sólo otro elemento, cualquiera que sea, entonces el otro elemento es la cabeza, ejemplo:

$$sn \leftarrow espec @grupnom$$
$$sn \leftarrow @grupnom espec$$

12. Para los patrones con frase nominal (*grupnom*) como nodo padre, si el patrón contiene un sustantivo (*n*) entonces se marca como cabeza, ejemplo:

$$grupnom \leftarrow s @n sp$$
$$grupnom \leftarrow @n sn$$
$$grupnom \leftarrow s @n S$$

13. Para los patrones con frase nominal (*grupnom*) como nodo padre, si el patrón contiene una frase nominal (*grupnom*), ésta se marca como cabeza, ejemplo:

$$grupnom \leftarrow @grupnom s$$
$$grupnom \leftarrow @grupnom sn$$

14. Para patrones con especificador (*espec*) como nodo padre, si el patrón contiene un artículo definido (*da*), entonces éste es la cabeza,

$$espec \leftarrow @da di$$
$$espec \leftarrow @da dn$$

15. Si el patrón contiene un adjetivo calificativo (*aq*) y una frase preposicional (*sp*), entonces el adjetivo es la cabeza:

$$grupnom \leftarrow @aq sp$$

16. Si el patrón contiene un *sadv* como primer elemento, seguido de un adjetivo calificativo (*aq*) y cualquier otro elemento, entonces el adjetivo es la cabeza:

$$S \leftarrow sadv @aq sp$$

El orden de aplicación de las reglas de heurística es importante. Por ejemplo, si aplicamos la regla 4 en el patrón: $S \leftarrow coord sn gv sn$, la cabeza sería la frase verbal (*gv*)

sin embargo el jefe de frase correcto debería ser el coordinante (coord). Por lo tanto la regla 2 debe ser aplicada primero.

Hay casos para los que no hay consenso en la comunidad de gramáticas de dependencia sobre la selección de cabezas (como coordinación, construcciones relativas, etc.). Las heurísticas anteriores reflejan solo una posible opción lingüística.

4.5 Evaluación de las heurísticas y corrección de la gramática

El algoritmo extrae 2668 reglas de gramática. De éstas, 346 (13%) se repiten mas de diez veces y el resto (2322, 87%) se repiten menos de diez veces. La figura 9 muestra las 20 reglas mas frecuentes con su respectivo número de ocurrencias.

Las heurísticas utilizadas marcan automáticamente los jefes de frase de 2226 (83%) de todas las reglas de gramática extraídas del corpus.

Para evaluarlas se seleccionaron aleatoriamente 300 reglas de las cubiertas por las heurísticas y se marcaron a mano. La comparación muestra que el 99% de las marcas de jefes de frase coinciden. La figura 10 muestra las reglas que no coinciden.

Los jefes de frase de estas reglas se corrigieron manualmente en la gramática extraída del corpus Cast3LB.

Considerando las estadísticas de comparación, creemos que al menos el 95% de las reglas del Cast3LB son marcadas correctamente.

4.6 Algoritmo para la transformación de un árbol de constituyentes en uno de dependencias

El algoritmo de transformación usa recursivamente la información de los patrones marcados con cabezas para determinar cuales componentes subirán en el árbol. Esto significa desconectar la cabeza de sus hermanos y colocarla en la posición del nodo padre.

```

13045 sn ← espec grupnom
12234 sp ← prep sn
3335 grupnom ← n sp
1902 grupnom ← n s
1224 sp ← prep S
996 grupnom ← n S
827 S ← S coord S
542 gv ← va vm
542 grupnom ← s n
540 S ← infinitiu sn
484 grupnom ← n s sp
430 sn ← sn coord sn
423 grupnom ← n sn
406 grupnom ← grupnom coord grupnom
402 S ← aq sp
377 grupnom ← s n sp
357 gv ← vm infinitiu
348 S ← sn gv sn
292 grupnom ← n sp sp
295 S ← S S

```

Figura 9. Reglas más frecuentes extraídas del corpus Cast3LB.

Marcadas automáticamente	Marcadas manualmente
infinitiu ← va vm sps00 @infinitiu	infinitiu ← va @vm sps00 infinitiu
S ← conj S @coord S	S.F.C.co-CD ← @conj S coord S

Figura 10. Reglas marcadas automáticamente no coinciden con las marcadas manualmente.

Para entenderlo más claramente, el algoritmo se detalla a continuación:

1. Recorremos el árbol de constituyentes en profundidad de izquierda a derecha, comenzando de la raíz y visitando sus nodos hijos recursivamente.
2. Para cada patrón en el árbol, buscamos en las reglas de la gramática para encontrar cuál elemento es la cabeza.
3. Marcamos la cabeza en el árbol de constituyentes.
4. Desconectamos la cabeza de sus hermanos y la colocamos en la posición del nodo padre.

El algoritmo termina cuando un nodo cabeza sube a la raíz del árbol. Para ilustrar un ejemplo consideramos las siguientes figuras.

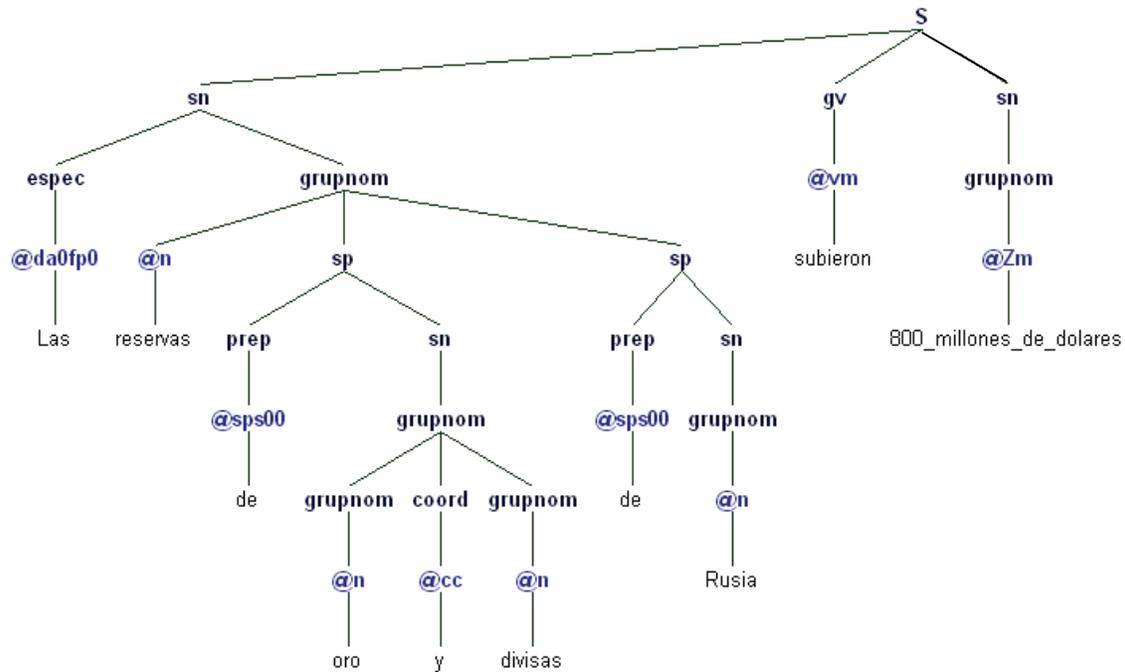


Figura 11. Árbol de constituyentes. 'Las reservas de oro y divisas de Rusia subieron 800_millones_de_dolares'

La figura 11 muestra un árbol de constituyentes que se transformará a un árbol de dependencias. También podemos observar que hay unos nodos marcados como cabezas, debido a que, como mencionamos anteriormente, los nodos que solo tienen una hoja se marcan en la extracción de la gramática.

Siguiendo el algoritmo, el primer patrón a encontrar (en el árbol de constituyentes de la figura 11) sería: $grupnom \leftarrow grupnom\ coord\ grupnom$.

Si buscamos en las reglas de gramática marcadas por las heurísticas encontramos que la cabeza de ese patrón es el coordinante (*coord*).

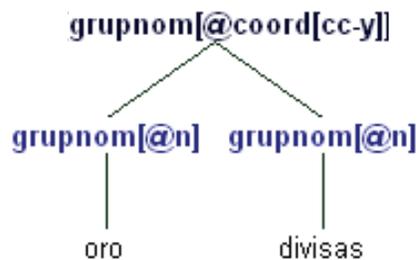


Figura 12. Nodos cabeza colocados en la posición del nodo padre del patrón.

Marcamos la cabeza en el árbol de constituyentes y la desconectamos poniéndola en la posición del nodo padre (del patrón al que pertenece). La figura 12 muestra este paso del algoritmo.

El algoritmo sigue su ejecución hasta que un nodo cabeza sube a la posición del nodo raíz.

4.6.1 El caso de los coordinantes

No existe consenso en los formalismos de dependencias de cómo manejar conjunciones coordinadas. Para nuestro caso se hizo lo siguiente:

Cuando se encuentra una coordinante (y, o) se duplica (o multiplica) la oración de forma tal que cada oración contendrá la relación sintáctica de un solo elemento. Por ejemplo, para la frase “nombres para niños y niñas” se interpreta como *nombres para niños* y *nombres para niñas*, creando así dos árboles diferentes.

Al hacer esto se aumenta la precisión del algoritmo, debido a que se pueden extraer árboles sintácticos más correctos y por lo tanto colocaciones más correctas.

Siguiendo con el ejemplo de la figura 11 y tomando en cuenta los casos de conjunciones coordinadas, el algoritmo extrae los árboles de dependencias con etiquetas mostrados en las figuras 13 y 14. En las figuras 15 y 16 se pueden observar los mismos árboles sin etiquetas para mayor claridad.

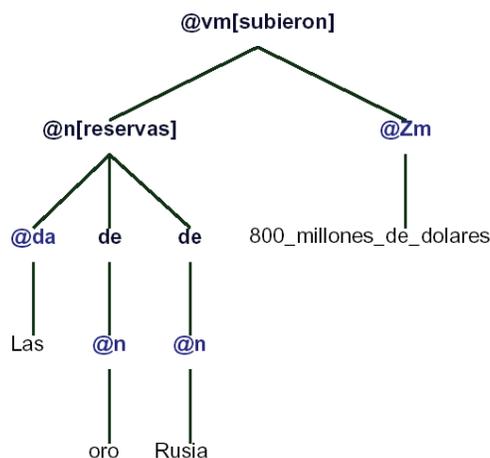


Figura 13. Primer árbol de dependencias resultante con etiquetas.

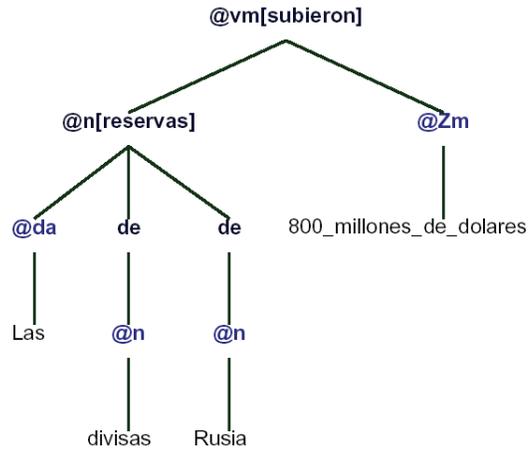


Figura 14. Segundo de dependencias resultante con etiquetas.

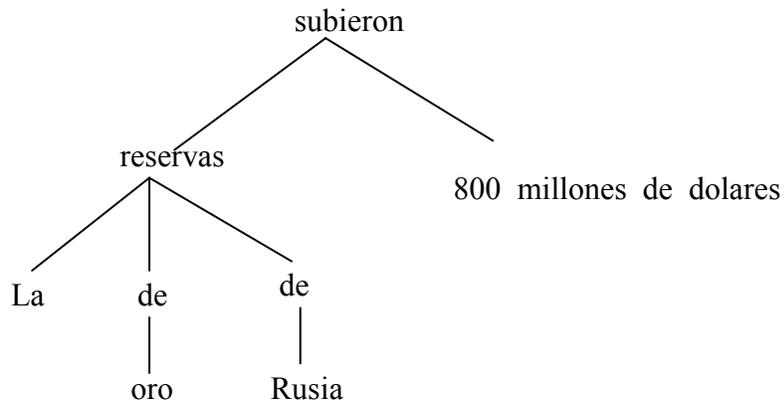


Figura 15. Primer árbol de dependencias sin etiquetas.

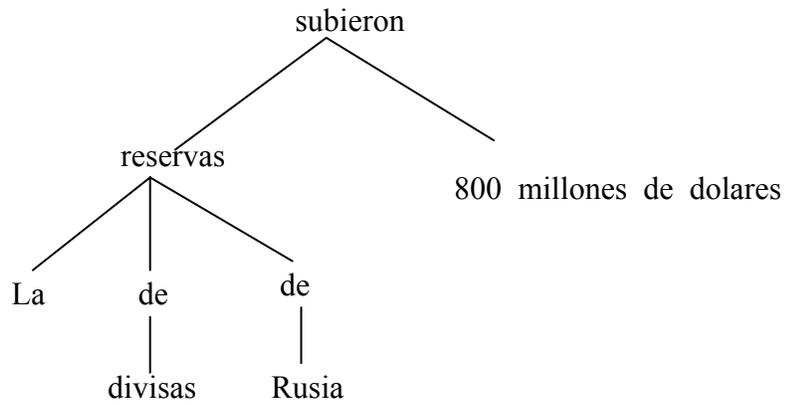


Figura 16. Segundo árbol de dependencias sin etiquetas.

4.7 Implementación

El algoritmo se implementó en módulos para disminuir el tiempo de búsqueda de errores y para generar resultados en cada módulo que permitieran seguir el proceso del algoritmo.

La mayoría de los módulos se programaron en Minimacro, y algunos otros en Perl, dos lenguajes de programación enfocados al procesamiento de texto, que hacen uso de expresiones regulares para hacer más fácil el manejo de cadenas.

La tabla siguiente muestra los módulos principales del algoritmo y una descripción la entrada y salida de cada módulo.

Tabla 1. Descripción de la implementación de los módulos principales del algoritmo de transformación.

Módulo	Descripción
<i>acentos.pl</i>	Cambia las vocales acentuadas a vocales sin acentos y entre guiones (ejemplo: <i>á</i> la cambia por <i>_a_</i>). Recibe el corpus Cast3LB y regresa el corpus sin acentos.
<i>clean.mm</i>	Se asegura que el paréntesis de inicio oración comience en una nueva línea. Recibe la salida de <i>acentos.pl</i> y regresa el corpus con inicio de oraciones en nueva línea.
<i>simpli.mm</i>	Simplifica el corpus quitando los datos adicionales de las etiquetas. Recibe la salida de <i>clean.mm</i> y regresa el corpus con etiquetas simplificadas.
<i>heurist.mm</i>	Este módulo marca las cabezas de los patrones extraídos del corpus utilizando reglas de heurística. Recibe la lista de patrones extraídos del corpus (salida del módulo <i>dependen.mm</i>) y regresa la misma lista con cabezas marcadas.
<i>dependen.mm</i>	Este módulo utiliza la salida de <i>heurist.mm</i> para marcar en el mismo corpus las cabezas de los patrones que va encontrando. Si no encuentra algún patrón lo agrega a una lista de patrones extraídos del corpus. Recibe la salida de <i>simpli.mm</i> y regresa el corpus marcado con las cabezas de cada patrón. También regresa una lista de patrones extraídos.
<i>coord.mm</i>	Este módulo duplica (o multiplica) la oración cuando encuentra un coordinante (y,o) poniendo cada elemento de la unión como cabeza en una oración diferente. Recibe la salida de <i>dependen.mm</i> y regresa un corpus de dependencias con oraciones que no contienen conjunciones coordinadas.
<i>relocate.mm</i>	Este modulo se encarga de desconectar las cabezas de cada patrón y colocarlas en la posición del nodo padre. Recibe la salida de <i>coord.mm</i> y regresa un corpus con las cabezas de patrones posicionadas en su nodo padre.

<i>convert.mm</i>	Este modulo elimina toda la información extra de los nodos padres dejando sólo la cabeza con su información léxica, convirtiendo así las oraciones en árboles de dependencias. Recibe la salida de <i>relocate.mm</i> y regresa un corpus de dependencias.
<i>desacentos.pl</i>	Este modulo busca las vocales que fueron desacentuadas para volverlas a acentuar. Recibe la salida de <i>convert.mm</i> y regresa un corpus de dependencias ya con acentos.

4.8 Metodología de evaluación

Seguimos el esquema de evaluación propuesto por [\[Briscoe et al, 2002\]](#), que sugiere evaluar la precisión en base a relaciones gramaticales entre cabezas lematizadas.

Para obtener los valores para las métricas de evaluación *precision* y *recall*, se extraen colocaciones de los árboles de dependencias resultantes de nuestro método y se compara con colocaciones extraídas manualmente del mismo corpus Cast3LB.

Consideramos la colocación como una relación entre un nodo padre con un nodo hijo y su tipo de relación. Por ejemplo, las colocaciones extraídas de la frase “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares” son:

reserva DET el
reserva de Rusia
reserva de oro
reserva de divisa
subir CIF 800_millones_de_dolares
subir SUST reserva

Para la evaluación, seleccionamos aleatoriamente 35 oraciones del corpus y se convirtieron manualmente a árboles de dependencias, de los cuales se extraen 419 colocaciones. Aplicamos nuestro algoritmo a esas mismas oraciones para obtener los árboles de dependencias y extraemos las colocaciones automáticamente.

4.9 Resultados obtenidos

El sistema extrae 417 colocaciones de los 35 árboles seleccionados aleatoriamente. De las 419 colocaciones extraídas manualmente, el sistema extrae 370 (88%) correctas.

De esta forma, los valores para las métricas de evaluación *precision* y *recall* de los árboles de dependencias extraídos del corpus Cast3LB son:

Métrica de evaluación	Valor
Precision	0.883
Recall	0.887

4.10 Conclusiones

La representación de dependencias de estructuras sintácticas tiene importantes ventajas en ciertas aplicaciones. Sin embargo, la mayoría de las herramientas y recursos léxicos existentes, especialmente los corpus con etiquetado sintáctico manual, son orientados al formalismo de constituyentes.

En esta sección se presentó una técnica para obtener automática o semiautomáticamente recursos sintácticos para el formalismo de dependencias, a través de los recursos existentes en el formalismo de constituyentes, proponiendo las heurísticas para la conversión de un corpus de constituyentes a un corpus de dependencias. El corpus obtenido permite la extracción de una gramática de dependencias.

CAPÍTULO

5

5

Extracción del diccionario de colocaciones

Capítulo 5. Extracción del diccionario de colocaciones

5.1 Introducción

Los lenguajes naturales están llenos de colocaciones, combinaciones recurrentes y arbitrarias de palabras que co-ocurren frecuentemente y no por casualidad. Las colocaciones son útiles en diferentes aplicaciones de procesamiento de lenguaje natural.

En este capítulo daremos una breve reseña de lo que son las colocaciones, sus propiedades, tipos de colocaciones y algunas aplicaciones para las que son útiles. También se presenta un algoritmo para la extracción del diccionario de colocaciones a partir de un corpus.

5.2 ¿Qué son las colocaciones?

Hay mucha discusión y trabajo relacionado con colocaciones [[Allerton, 84](#); [Cruse, 86](#); [Mel'cuk, 81](#)]. 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 de Benson [[Benson, 90](#)]. La definición es la siguiente: Una colocación es una combinación recurrente y arbitraria de palabras.

Éstas cubren pares de palabras y frases que son comúnmente usadas en el lenguaje para las cuales no se aplican reglas semánticas o sintácticas.

5.3 Propiedades de las colocaciones

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

5.3.1 Las colocaciones son arbitrarias

Las colocaciones son difíciles de producir para un hablante no nativo [Nakhimovsky & Leed 79]. 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 2 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.

Tabla 2. 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	enfoncez 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

La co-ocurrencia 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".

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

5.3.2 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 [[Dellenbaugh & Dellenbaugh 90](#)], 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 técnica, 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.

5.3.3 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.

5.3.4 Las colocaciones son conjuntos de cohesión léxica

Por cohesión léxica [[Smadja, 93](#)] 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 [[Cowie, 81](#); [Benson, 89a](#)].

Los lexicógrafos usan el juicio lingüístico de la gente para decir cuales son colocaciones y cuales no. Ellos aplican cuestionarios a la gente, como el que se muestra en la figura 17.

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"

Figura 17. Test llenar-el-espacio, de [\[Benson, 90\]](#).

Este cuestionario contiene las oraciones usadas por Benson para compilar el conocimiento de colocaciones para el diccionario BBI [\[Benson, 89b\]](#). 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 está dadas en la segunda columna.

Como consecuencia, las colocaciones tienen una distribución estadística particular [\[Halliday 66; Cruse 86\]](#). 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.

5.4 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 [[Smadja, 93](#)]: oraciones nominales rígidas, relaciones predicativas y plantillas de frase. A continuación se explican cada una de ellas.

5.4.1 Relaciones predicativas

Una relación predicativa consiste en dos palabras que se usan juntas repetidamente en una relación sintáctica similar [[Smadja, 93](#)]. 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 [[Mel'cuk, 81](#)], y las relaciones tipo L de Benson [[Benson et al, 86b](#)].

5.4.2 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 [[Choueka, 88](#)] y [[Amsler, 89](#)]. 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.

5.4.3 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.

5.5 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:

- Desambiguación
- Traducción automática
- Generación de lenguaje
- Recuperación de Información
- Lexicografía computacional
- Lingüística basada en corpus e investigación sociológica

5.6 Algoritmo para la extracción del diccionario de colocaciones del corpus

En el capítulo anterior presentamos el algoritmo y su implementación para la transformación del corpus de constituyentes Cast3LB en un corpus de dependencias. De esta forma, una vez que tenemos el corpus de dependencias aplicamos un algoritmo para extraer las colocaciones, el cual se describe a continuación.

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.

Para ilustrar un ejemplo consideramos la siguiente oración: “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares” (los árboles de dependencias extraídos de esta oración se muestra en las figuras 15 y 16)¹.

Recorriendo el primer árbol(figura 15), visitamos el primer nodo que sería la raíz y encontramos que la primer colocación a extraer es *subieron SUST reservas*, donde *subieron* es el nodo padre, *reservas* es el nodo hijo y *SUST* es la relación de dependencia entre ellos. Siguiendo el algoritmo, recorreremos el árbol en profundidad de izquierda a derecha y visitamos el siguiente nodo que sería *reservas* y la siguiente colocación a extraer sería *reservas DET Las*. Y así sucesivamente hasta recorrer por completo el árbol. En la figura 18 se muestran encerradas en óvalos las colocaciones a extraer del primer árbol de dependencias de la oración. El número que aparece al lado de cada colocación es el orden en el que se extraen.

Las colocaciones extraídas automáticamente de la oración “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares” se muestran en la figura 19.

¹ Se extrajeron dos árboles de dependencias debido al coordinante *y* que contiene ésta oración, ver punto 4.6.1

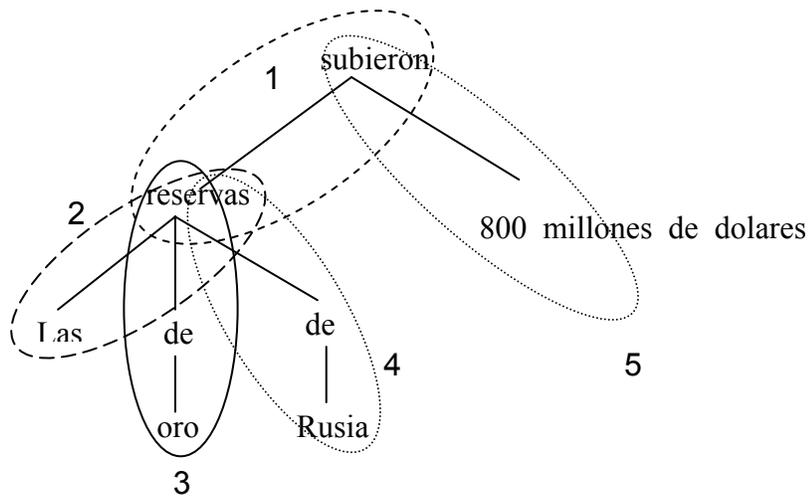


Figura 18. Colocaciones a extraer del primer árbol de dependencias de la oración “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares”.

subieron SUST reservas
 reservas DET Las
 reservas de oro
 reservas de divisas
 reservas de Rusia
 subieron CIF 800_millones_de_dolares

Figura 19. Colocaciones extraídas automáticamente de la oración “Las reservas de oro y divisas de Rusia subieron 800_millones_de_dólares”.

Tomando en cuenta la idea presentada anteriormente de resolver la ambigüedad sintáctica, agregamos información estadística (frecuencias) a las colocaciones extraídas.

5.7 Implementación

El algoritmo se implementó en módulos para disminuir el tiempo de búsqueda de errores y para generar resultados en cada módulo que permitieran seguir el proceso del algoritmo.

La mayoría de los módulos se programaron en Minimacro, y algunos otros en Perl, dos lenguajes de programación enfocados al procesamiento de texto, que hacen uso de expresiones regulares para hacer más fácil el manejo de cadenas.

La tabla siguiente muestra los módulos principales del algoritmo y una descripción la entrada y salida de cada módulo.

Tabla 3. Descripción de la implementación de los módulos principales del algoritmo de extracción.

Módulo	Descripción
<i>acentos.pl</i>	Cambia las vocales acentuadas a vocales sin acentos y entre guiones (ejemplo: <i>á</i> la cambia por <i>_a_</i>). Recibe el corpus de dependencias y regresa el mismo corpus sin acentos.
<i>treeview.mm</i>	Este módulo toma las oraciones de dependencias y las pone en forma de árbol, especificando niveles por medio de sangrías. Recibe la salida de <i>acentos.pl</i> y regresa el corpus en forma de árboles.
<i>tree2par.mm</i>	Extrae las relaciones entre nodos padres e hijos utilizando la estructura de forma de árbol. Recibe la salida de <i>treeview.mm</i> y regresa una lista de relaciones entre padres e hijos.
<i>goodpair.mm</i>	Extrae las colocaciones. Recibe la salida de <i>tree2par.mm</i> y regresa una lista de colocaciones.
<i>desacentos.pl</i>	Este modulo busca las vocales que fueron desacentuadas para volverlas a acentuar. Recibe la salida de <i>goodpair.mm</i> y regresa la lista de colocaciones ya con acentos.

Para agregar las frecuencias de cada colocación no implementamos ningún módulo, ya que se utilizaron dos programas ya existentes; *lsort.exe* que ordena alfabéticamente las colocaciones y *unique.exe* que cuenta pone las frecuencias de cada colocación y elimina las repetidas.

5.8 Resultados obtenidos

El algoritmo extrae automáticamente 46,698 colocaciones del corpus Cast3LB.

Para evaluación, seleccionamos aleatoriamente 35 oraciones del corpus y se extrajeron manualmente 419 colocaciones. Aplicamos nuestro algoritmo a esas mismas oraciones obteniendo automáticamente 417 colocaciones, de las cuales 370 (88%) coinciden con las extraídas manualmente.

Considerando estas estadísticas, inferimos que más del 85% (aproximadamente 41,000) de las colocaciones extraídas automáticamente del corpus Cast3LB son correctas. Además, las colocaciones extraídas contienen información estadística (frecuencias), obtenida automáticamente a partir del corpus.

5.9 Conclusiones

En esta sección se presentó una técnica para obtener automáticamente colocaciones a partir de un corpus de constituyentes. Se obtuvo una base de datos con más de 40,000 colocaciones.

Como mencionamos anteriormente, las colocaciones son útiles en diferentes aplicaciones de procesamiento de lenguaje natural. Una de ellas resolver la ambigüedad sintáctica que es uno de los problemas más difíciles que se presentan en sistemas de procesamiento de lenguaje natural.

Otra aplicación importante es la utilización de colocaciones para la evaluación de analizadores sintácticos de dependencias.

CAPÍTULO

6

Evaluación del parser de dependencias DILUCT

Capítulo 6. Evaluación del parser de dependencias DILUCT

6.1 Introducción

Un parser de constituyentes estándar construye una estructura incremental, de modo que la falta de construcción de un componente implica la imposibilidad de construir todos los demás componentes que contengan a éste. Lo que significa que una decisión incorrecta sobre un primer paso del análisis del parser conduce a un resultado final totalmente o en gran parte incorrecto.

En cambio, en el parser de dependencias la selección de un rector de una palabra dada, o la selección de “existencia” si dos palabras dadas están conectadas con una relación de dependencia (relación sintáctica), es independiente de la decisión correspondiente de otro par de palabras. Esto permite continuar el proceso de análisis aunque algunas decisiones no se tomen con éxito. La estructura resultante puede ser incompleta (faltándole algunas relaciones) o no correcta totalmente (con algunas relaciones identificadas incorrectas). Sin embargo, una decisión incorrecta sobre un par particular de palabras generalmente no causa una serie de errores conectados en los pasos posteriores del análisis.

En este capítulo describimos el funcionamiento del parser de dependencias para el español DILUCT, que utiliza un conjunto ordenado de reglas de heurísticas simples para determinar de forma iterativa las relaciones de dependencia entre palabras todavía no asignadas a un rector. En caso de ambigüedades de ciertos tipos, las estadísticas de co-ocurrencia de la palabra buscada de manera no-supervisado en un corpus grande o en la Web se utilizan para seleccionar la variante más probable.

Después se presenta la metodología que se utilizó para evaluar el parser DILUCT con el diccionario de colocaciones extraído del corpus Cast3LB, así como los resultados obtenidos.

6.2 Funcionamiento del parser DILUCT²

Siguiendo la aproximación estándar, el parser primero preprocesa el texto de entrada – básicamente incluye tokenización, separación de oraciones, etiquetado de categorías gramaticales (en inglés, part of speech, POS) y lematización- y luego aplica el algoritmo de análisis (parsing).

6.2.1 Pre-procesamiento

Tokenización y separación de oraciones

El texto es tokenizado en palabras y signos de puntuación, y separado en oraciones.

Actualmente no se distingue entre signos de puntuación, es decir, cualquier signo de puntuación se sustituye con una coma.

Dos componentes de artículo y preposición se separan: *del* → *de el*, *al* → *a el*.

Las preposiciones compuestas representadas en la escritura como palabras separadas se unen en una sola palabra, ejemplo: *con la intención de*, *a lo largo de*, etc. De la misma manera se tratan unas cuantas frases adverbiales como *a pesar de*, *de otra manera*, etc., y algunas frases pronominales como *sí mismo*.. La lista de tales combinaciones es pequeña (incluye 62 combinaciones) y cerrada. Actualmente no se lleva a cabo reconocimiento de nombres de entidades pero se planea para el futuro.

Etiquetado

El texto es etiquetado con categorías gramaticales (POS-tagged) usando el etiquetador TnT [[Brants, 2000](#)] entrenado en el corpus de español CLiC-TALP. Este etiquetador tiene 94% de exactitud en su funcionamiento [[Morales-Carrasco & Gelbukh, 2003](#)].

² Tomado de [[Calvo and Gelbukh, 2006](#)]

Además se corrigen algunos errores frecuentes del etiquetador TnT, por ejemplo:

Regla	Ejemplo
Det Adj V → Det S V	<i>el inglés vino</i>
Det Adj Prep → Det S Prep	<i>el inglés con</i>

Lematización

Se usa un diccionario basado en un analizador morfológico para español [Gelbukh *et al.*, 2003]. En caso de ambigüedad, la variante de la categoría gramatical (POS) elegida por el etiquetador es seleccionada, con las siguientes excepciones:

Etiquetado por el tagger	Encontrado por el analizador	Ejemplo
Adjetivo	Pasado participio	<i>dado</i>
Adverbio	Presente participio	<i>dando</i>
Sustantivo	Infinitivo	<i>dar</i>

Si el analizador no da una opción en la primera columna pero sí en la segunda, éste último es aceptado.

Si un sustantivo, adjetivo o participio no es reconocido por el analizador, se eliminan los sufijos, ejemplo: *flaquito* → *flaco*. Para esto, se trata de eliminar sufijos esperados y checar si la palabra es reconocida por el analizador morfológico. Ejemplos de reglas de sufijos a eliminar son:

Regla	Ejemplo
<i>-cita</i> → <i>-za</i>	<i>tacita</i> → <i>taza</i>
<i>-quilla</i> → <i>-ca</i>	<i>chiquilla</i> → <i>chica</i>

6.2.2 Reglas

Las reglas de análisis (parsing) se aplican al texto ya lematizado. Siguiendo una aproximación similar a [Apresyan *et al.*, 89; Calvo & Gelbukh, 2003], se representa una regla como un sub-grafo, es decir, $N \leftarrow V$. La aplicación de una regla consiste en los siguientes pasos:

1. Una subcadena que coincide con la secuencia de palabras en la regla se busca en la oración.
2. Las relaciones sintácticas entre las palabras que coincidieron se establecen de acuerdo a las especificaciones de esa regla.

3. Todas las palabras que han sido asignadas a un rector por la regla son removidas de la oración en el sentido de que no participan en futuras comparaciones del paso 1.

Por ejemplo, para la oración *Un perro grande ladra*:

Oración	Regla
<i>Un(Det) perro(N) grande(Adj) ladra (V)</i>	Det ← N
<i>perro(N) grande(Adj) ladra (V)</i> ↓ <i>Un(Det)</i>	N → Adj
<i>perro(N) ladra (V)</i> ↙ ↘ <i>Un(Det) grande(Adj)</i>	N ← V
<i>ladra (V)</i> ↓ <i>perro(N)</i> ↙ ↘ <i>Un(Det) grande(Adj)</i>	Hecho

Como puede verse en el ejemplo, el orden de aplicación de la regla es importante. Las reglas son ordenadas, es decir, para cada iteración del algoritmo, la primera regla que se pueda aplicar se aplica, y entonces el algoritmo repite buscando reglas aplicables para el primero. El proceso termina cuando no existen más reglas por ser aplicadas.

Note que una de las consecuencias de tal algoritmo es el tratamiento natural de modificadores repetidos. Por ejemplo, en las frases *el otro día* o *libro nuevo interesante* los dos determinantes (dos adjetivos respectivamente) serán conectados como modificadores del sustantivo por la misma regla Det ← N (N → Adj, respectivamente) en dos iteraciones sucesivas del algoritmo.

Las reglas todavía no se formalizan completamente (por eso su aproximación se llama semi-heurística), por ello más adelante se dan comentarios adicionales a algunas reglas.

Actualmente la gramática incluye las siguientes reglas:

Regla	Ejemplo
Verbos auxiliares y cadenas de verbos	
$estar \mid andar \leftarrow Ger$	<i>estar comiendo</i>
$haber \mid ser \leftarrow Part$	<i>haber comido</i>
$haber \leftarrow estado \leftarrow Ger$	<i>haber estado comiendo</i>
$ir_{pres} a \leftarrow Inf$	<i>ir a comer</i>
$ir_{pres} \leftarrow Ger \leftarrow Inf$	<i>ir queriendo comer</i>
$V \rightarrow que \rightarrow Inf$	<i>tener que comer</i>
$V \rightarrow V$	<i>querer comer</i>
Construcciones estándar	
$Adv \leftarrow Adj$	<i>muy alegre</i>
$Det \leftarrow N$	<i>un hombre</i>
$N \rightarrow Adj$	<i>hombre alto</i>
$Adj \leftarrow N$	<i>gran hombre</i>
$V \rightarrow Adv$	<i>venir tarde</i>
$Adv \leftarrow V$	<i>perfectamente entender</i>
Conjunciones (ver la explicación siguiente)	
$N \text{ Conj } N \text{ V(pl)} \Rightarrow [N N] \text{ V(pl)}$	<i>Juan y María hablan</i>
$X \text{ Conj } X \Rightarrow [X X]$	<i>(libro) nuevo e interesante</i>
(X se aplica a cualquiera)	
Otras reglas	
$N \rightarrow que V$	<i>hombre que habla</i>
$que \rightarrow V$	<i>que habla</i>
$\left\{ \begin{array}{l} \downarrow \\ \downarrow \end{array} \right.$	<i>hombre tal que; hombre , que</i>
$N X que$	
(X se aplica a cualquiera)	
$Det \leftarrow Pron$	<i>otro yo</i>
$V \rightarrow Adj$	<i>sentir triste</i>
$\left\{ \begin{array}{l} \downarrow \\ \downarrow \end{array} \right.$	<i>hombre , alto</i>
N , Adj	
$\left\{ \begin{array}{l} \downarrow \\ \downarrow \end{array} \right.$	<i>hombre , mujer</i>
N , N	
$N \rightarrow Prep \rightarrow V$	<i>obligación de hablar</i>
$\left\{ \begin{array}{l} \downarrow \\ \downarrow \end{array} \right.$	<i>comer , dormir</i>
V , V	
$V Det \leftarrow V$	<i>aborrecer el hacer</i>

Las conjunciones coordinadas siempre han sido el “dolor de cabeza” de los formalismos de dependencias y un argumento a favor de los de constituyentes. Siguiendo la idea de [Gladki, 85], se representan las palabras coordinadas de la forma como lo hacen los constituyentes, uniéndolas en una cuasi-palabra compuesta. En el árbol resultante se duplica (o multiplica) cada arco que viene o sale de un nodo especial. Por ejemplo, [Juan María] \leftarrow hablan (Juan y María hablan) es interpretado como la representación de dos relaciones: Juan \leftarrow habla y María \leftarrow habla.

Por consiguiente, las reglas para manejar conjunciones se reescriben como nuevas reglas más que como reglas de construcción de árbol. La primer regla forma una cuasi-palabra compuesta de dos sustantivos coordinados si les precede un verbo en plural. La regla elimina la conjunción, así la implementación de conjunciones no participa en la estructura del árbol.

6.2.3 Asignación de frase preposicional

Esta fase es llevada a cabo después de la aplicación de reglas descrita anteriormente.

Para cualquier preposición que no ha sido asignada a un rector, su compatibilidad con cualquier sustantivo y cualquier verbo en la oración es evaluada usando estadísticas de co-ocurrencia de la palabra (que se puede obtener por una simple búsqueda en Internet). La medida obtenida se combina con una penalización en la distancia lineal: El más distante es un rector potencial para la preposición en la pregunta que menos apropiada es para su asignación.

Los detalles de esta técnica estadística para el uso de preposiciones se encuentra en [\[Calvo & Gelbukh, 2004\]](#).

6.2.4 Heurísticas

El propósito de las heurísticas es asignar un rector a las palabras que no fueron asignadas a ninguno en el estado de aplicación de reglas.

El sistema actualmente usa las siguientes heurísticas, las cuales son aplicadas en iteraciones en este orden, de forma similar a como se aplican las reglas:

1. Un *que* sin asignar es asignado al verbo mas cercano (a la izquierda o a la derecha del *que*) que no tiene otro *que* como su rector inmediato o indirecto.
2. Para un pronombre no asignado se asigna al verbo más cercano que no tiene un *que* como su rector inmediato o indirecto.
3. Un sustantivo sin asignar es asignado al verbo más probable que no contiene un *que* como su rector inmediato o indirecto. Para estimar la probabilidad, un

algoritmo similar a el que está descrito en la sección anterior se utiliza. La estadística descrita en [Calvo *et al.*, 2005] se utiliza.

4. Para un verbo v no asignado aún, se busca a la izquierda el verbo más cercano w ; si no hay verbos a la izquierda, entonces se busca el más cercano a la derecha. Si w tiene un *que* como rector directo o indirecto, entonces v es asignado al *que*; de lo contrario se asigna a w .
5. Un adverbio o conjunción subordinada sin asignación (excepto *que*) es asignado al verbo más cercano (a la izquierda o a la derecha del *que*) que no tiene otro *que* como su rector inmediato o indirecto.

Note que si la oración contiene más de un verbo, en el paso 4 cada verbo es asignado a algún otro verbo, lo cual puede resultar en una dependencia circular. De cualquier forma, esto no daña puesto que una dependencia circular se romperá en el último paso del procesamiento.

6.2.5 Selección de la raíz

La estructura construida en los pasos anteriores del algoritmo descrito puede ser redundante. Particularmente, puede contener dependencias circulares entre verbos. El paso final del análisis es seleccionar la raíz más apropiada.

Se usa las siguientes heurísticas para seleccionar la raíz. Para cada nodo en el digrafo obtenido, contamos el número de otros nodos accesibles dado a través de una trayectoria dirigida a lo largo de las flechas. La palabra que maximiza este número es seleccionada como raíz. Particularmente, todos los arcos entrantes se eliminan de la estructura final.

6.3 Algoritmo para la evaluación del parser DILUCT con el corpus Cast3LB

En esta sección se presenta una comparación del parser DILUCT contra el corpus Cast3LB. Además se compara el parser DILUCT con dos parsers conocidos para el

español. El primero es Connexor, un parser de dependencias y el segundo TACAT, un parser de constituyentes.

Se siguió el esquema de evaluación propuesto por [\[Briscoe et al, 2002\]](#), que sugiere evaluar la precisión en base a relaciones gramaticales entre cabezas lematizadas. Este esquema es adecuado para evaluar parsers de dependencias y algunos de constituyentes, porque considera relaciones en un árbol, las cuales se presentan en ambos formalismos.

Para la evaluación se convirtió la salida de los árboles analizados por DILUCT y Connexor en colocaciones para compararlas con las colocaciones del corpus Cast3LB, extraídas en esta tesis.

6.4 Resultados obtenidos

Se seleccionaron aleatoriamente 190 oraciones del corpus Cast3LB y se analizaron con Connexor y DILUCT. A continuación se muestran los valores para las métricas de evaluación *precision* y *recall* de los diferentes parsers contra las colocaciones del corpus Cast3LB.

	Precision	Recall
Connexor	0.55	0.38
DILUCT	0.47	0.55

6.5 Conclusiones

En este capítulo dimos una breve descripción del funcionamiento del parser DILUCT, un parser de dependencias para el español.

Se describió la metodología para la evaluación de un analizador sintáctico de dependencias, en este caso DILUCT, donde se utiliza el diccionario de colocaciones extraído en esta tesis.

CAPÍTULO

7

Conclusiones y trabajo futuro

Capítulo 7. Conclusiones y trabajo futuro

7.1 Discusión

Los recursos léxicos existentes, especialmente los corpus con etiquetado sintáctico manual, en su mayoría son orientados al formalismo de constituyentes –en parte por la inercia de las escuelas tradicionales y en parte porque resultan de proyectos de largo plazo cuyo desarrollo se empezó hace varios años.

La representación de dependencias simplifica grandemente algunas tareas con respecto a la de constituyentes. Por ejemplo:

- En lexicografía, recopilar estadísticas sobre la combinación sintáctica de palabras individuales.
- En recuperación de información y minería del texto, encontrar una frase o hacer una búsqueda compleja en un corpus.
- En el análisis semántico, transformando el árbol de dependencias en cualquier representación semántica más cercana.

De ahí surge la necesidad de la obtención automática o semiautomática de los recursos sintácticos para el formalismo de dependencias utilizando los recursos existentes en el formalismo de constituyentes.

7.2 Conclusiones

En base al trabajo desarrollado en esta tesis podemos concluir que:

Se desarrolló e implementó un algoritmo para construir árboles sintácticos de dependencias, a partir de un corpus de constituyentes. En resumen se describe el procedimiento como sigue:

- Extraer una gramática de gran escala a partir de un corpus de constituyentes.
- Desarrollar y aplicar las heurísticas para determinar los jefes de frase de cada patrón extraído.
- Marcar las cabezas dentro del corpus.
- Desconectar nodos cabeza y posicionarlos en el lugar del nodo padre hasta terminar en la raíz.

La evaluación del algoritmo muestra una precisión de 0.883 y un *racall* de 0.887.

Se desarrolló e implementó un algoritmo para extraer colocaciones sintácticas de un corpus etiquetado con constituyentes. Se describe en resumen el algoritmo como sigue:

- A partir de los árboles de dependencias construidos, se recorre el árbol en profundidad de izquierda a derecha.
- 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.
- Se agregaron las frecuencias de cada colocación.

Se extrajeron 46,698 colocaciones del corpus de constituyentes de las cuales más del 85% (aproximadamente 41,000) son correctas.

Se aplicaron los recursos obtenidos (diccionario de colocaciones) para la evaluación del parser de dependencias para el español, DILUCT.

7.3 Aportaciones

La mayoría de los recursos léxicos existentes son orientados al formalismo de constituyentes. Estos recursos siguen creándose hoy en día.

Tener un método para transformar automáticamente corpus de constituyentes a dependencias tiene ventajas debido a que: cada corpus se limita al trabajo elaborado por un grupo de personas específicas que utilizan ciertos criterios, unir dos corpus de constituyentes creados por diferentes personas es difícil, mientras que un algoritmo de transformación puede tomar varios corpus de constituyentes para crear nuevos corpus.

Más específicamente, las aportaciones que se derivan de este trabajo se han dividido en dos rubros: aportaciones al conocimiento y aportaciones técnicas, las cuales se describen a continuación.

7.3.1 Aportaciones al conocimiento

- Desarrollo, implementación y evaluación de un algoritmo para la conversión de un corpus con estructuras sintáctica etiquetadas con de constituyentes al corpus de dependencias.
- Desarrollo, implementación y evaluación de las heurísticas para la extracción, a partir de un corpus de constituyentes, de una gramática de gran escala capaz de construir los árboles sintácticos de dependencia.
- Desarrollo, implementación y evaluación de un algoritmo para la extracción de colocaciones sintácticas de un corpus etiquetado con constituyentes, incluido el tratamiento adecuado de preposiciones y conjunciones.
- Implementación de los resultados obtenidos en la evaluación de un analizador sintáctico de dependencias.

7.3.2 Aportaciones técnicas

- **Corpus** de español etiquetado con las **dependencias** sintácticas, obtenido automáticamente a través de conversión de un corpus de constituyentes.

- **Gramática** de español a gran escala, capaz de generar los árboles de dependencias, obtenida automáticamente a partir del corpus.
- **Base de datos** de colocaciones sintácticas de español, con información **estadística**, obtenida automáticamente a partir del corpus.

7.4 Publicaciones generadas

Alexander Gelbukh, Hiram Calvo, Sulema Torres. *Transforming a Constituency Treebank into a Dependency Treebank*. Procesamiento de Lenguaje Natural, No. 35, ISSN 1135-5948, España.

7.5 Trabajo futuro

- Utilizar el diccionario de colocaciones extraído para resolver ambigüedad sintáctica, siguiendo la idea presentada en esta tesis.
- Investigar cuáles parsers de dependencia se entrenan con gramática para integrarles la gramática extraída y hacerlos más robustos.
- Usar el diccionario de colocaciones extraído en desambiguación de sentidos de palabra (WSD) y ver los resultados.
- Hacer mejoras a los algoritmos presentados anteriormente 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).

Glosario

Glosario

Ambigüedad	Término que hace referencia a aquellas estructuras gramaticales que pueden entenderse de varios modos o admitir distintas interpretaciones y dar, por consiguiente, motivo a dudas, incertidumbre o confusión.
Ambigüedad léxica	La ambigüedad léxica es aquella que se presenta en la categoría gramatical de un vocablo. Es decir, un vocablo puede tener más de un rol gramatical en diferentes contextos.
Ambigüedad semántica	La ambigüedad semántica es aquella que se presenta en una estructura gramatical, de tal manera que ésta puede expresar diferentes sentidos dependiendo del contexto local, el tópico global y el mundo pragmático en el que se manifiesta.
Ambigüedad sintáctica	La ambigüedad sintáctica, también conocida como estructural, es aquella que se presenta en oraciones, de tal manera que éstas puedan ser representadas por más de una estructura sintáctica.
Analizador sintáctico	Un analizador sintáctico para un lenguaje natural, es un programa que construye árboles de estructura de frase o de derivación para las oraciones de dicho lenguaje, además de proporcionar un análisis gramatical, separando los términos en constituyentes y etiquetando cada uno de ellos. Asimismo, puede proporcionar información adicional acerca de las clases semánticas (persona, género) de cada palabra y también la clase funcional (sujeto, objeto directo, etc.) de los constituyentes de la oración.
Aprendizaje no supervisado	En el aprendizaje no supervisado no existe un profesor que corrija los errores al alumno; ya que éste recuerda más al autoaprendizaje. El alumno dispone del material de estudio; pero nadie lo controla. En el caso de la desambiguación no supervisada también es posible usar un corpus de texto etiquetado como fase entrenamiento; sin embargo los algoritmos no supervisados generalizan esta información para cualquier vocablo ambiguo; aunque no haya estado presente en

Aprendizaje supervisado	<p>el corpus de entrenamiento.</p> <p>El aprendizaje supervisado se asemeja al método de enseñanza tradicional con un profesor que indica y corrige los errores del alumno hasta que éste aprende la lección. En el caso de la desambiguación supervisada se entrena un clasificador usando un corpus de texto etiquetado semánticamente para obtener el contexto en el que usualmente se presenta cada sentido del vocablo ambiguo. Este clasificador desmabiguará solo aquellos vocablos y sentidos que hayan participado en el entrenamiento previo.</p>
Árbol de constituyentes	<p>Un árbol de constituyentes es una estructura de datos que pemite categorizar una oración en sus partes de oración.En el llamado sistema o método de constituyentes la principal operación lógica es la inclusión de elementos en conjuntos, así éstos pertenecen a una oración o a una categoría. Según esta aproximación, una oración es segmentada en constituyentes, cada uno de los cuales es consecuentemente segmentado. Así, esto favorece un punto de vista analítico.</p>
Árbol de dependencias	<p>Un árbol de dependencias es una estructura de datos que pemite obtener las relaciones de dependencia sintáctica entre un núcleo y conjunto de modificadores. La aproximación de dependencias se centra en las relaciones entre las unidades sintácticas últimas, es decir, en las palabras. La principal operación aquí consiste en establecer relaciones binarias. Según esta idea, una oración se construye de palabras, unidas por dependencias.</p>
Cabeza	<p>Es un término utilizado en este trabajo para referenciar al vocablo que gobierna una relación de dependencia sintáctica, de tal manera que se se puede obtener muchos modificadores sintácticos para una misma cabeza.</p>
Categoría gramatical	<p>El término categoría gramatical o parte de la oración, que en inglés se denomina POS (part of speech) es una clasificación de las palabras de acuerdo a la función que desempeñan en la oración. La gramática tradicional distingue nueve categorías gramaticales: sustantivo, determinante, adjetivo, pronombre, preposición, conjunción, verbo, adverbio, interjección. No obstante, para algunos lingüistas, las categorías gramaticales son una forma de clasificar ciertos rasgos gramaticales, como por ejemplo: modo, aspecto, tiempo y voz.</p>

Corpus / córpora	Un corpus es una recopilación de textos producidos en situaciones reales (" <i>pieces of language</i> ") y la inclusión de los textos que componen el corpus está guiada por una serie de criterios lingüísticos explícitos para asegurar que pueda usarse como muestra representativa de una lengua y facilitar su uso en tareas de procesamiento de lenguaje natural.
Criterios Lingüísticos	Se refiere a etiquetado de corpus, ya sea sintáctico, semántico, etc.
Dominio	El término dominio hace referencia a la temática general que expresa un documento o texto en su totalidad.
Etiqueta sintáctica	Se refiere a la combinación de letras y números que se agrega como información a una palabra para identificar su categoría gramatical.
Fonología	La fonología es el estudio de los sonidos del lenguaje. La fonética es la parte de la fonología que trata de la manera en que se pronuncian los sonidos y su forma acústica, pero la fonología también incluye el estudio de la manera en que los sonidos funcionan sistemáticamente en la lengua. Las otras divisiones importantes de ese estudio incluyen el análisis de los fonemas, los cambios de un sonido a otro en ciertos contextos (la alternancia morfofonémica y alofónica), la estructura de las sílabas, el acento, la entonación, y el tono lingüístico. [Inglés: phonology]
Frase	Grupo de una o más palabras que funciona como unidad, pero que (normalmente) no funciona en su totalidad independientemente, como en el caso de una oración. A veces se marcan las frases, como las oraciones, poniéndolas entre corchetes. Por ejemplo: [las montañas [más altas]] es una frase nominal, y también contiene una frase adjetival, [más altas]. Contrástese con oración. [Inglés: phrase]
Gramática	Es la manera característica en que se combinan los elementos básicos (especialmente los elementos léxicos) de una lengua, para formar estructuras más complejas que permitan la comunicación de los pensamientos. La gramática incluye la morfología y la sintaxis; algunos analistas incluyen la fonología, la semántica y el léxico también como parte de la gramática. [Inglés: grammar]
Herramientas lingüísticas	Esta expresión hace referencia a diversos programas o aplicaciones usados en el procesamiento de lenguaje natural, tales como analizadores sintácticos, morfológicos, corpus, diccionarios electrónicos, ontologías, entre otros.

H-nucleo	De aquí en adelante h-núcleo representa el término en inglés head. En la gramática tradicional se utiliza el término núcleo para las palabras o grupos de palabras más importantes. En la literatura de constituyentes head es el constituyente más importante gramaticalmente. Por ejemplo, en un grupo nominal el sustantivo es el head o núcleo. Sin embargo en la asignación de head a la frase completa difieren diferentes formalismos por lo que optamos por esta convención.
Inteligencia Artificial (IA)	Disciplina dedicada a desarrollar y aplicar enfoques computacionales al comportamiento inteligente. Estudia preferentemente los comportamientos y fenómenos de percepción, solución de problemas, razonamiento, utilización de un lenguaje natural y planeamiento de actividades. [Inglés: Artificial Intelligence (AI)]
Lemma, Lematización	Término en Latin. Forma canónica. En lexicografía, entrada léxica del diccionario en que se suministra diversa información y es a menudo representativa de distintas formas flexionadas; p. ej. ir es el lema de voy, vas, íbamos, fueron y el resto de sus formas conjugadas. Lematización. En lexicografía se denomina ‘lematización’ el proceso de reducción de las diferentes formas flexivas de una palabra a la forma canónica que se selecciona como lemma. [Inglés: lemma]
Lenguaje natural	Es un término ya adoptado que el lenguaje humano se denomine natural para diferenciarlo de los lenguajes artificiales en el área de la computación.
Lexema	Unidad léxica abstracta que no puede descomponerse en otras menores, aunque sí combinarse con otras para formar compuestos, y que posee un significado definible por el diccionario, no por la gramática. Por ejemplo: fácil es el lexema básico de facilidad, facilitar, fácilmente. [Inglés: lexeme]
Léxico	El conjunto de los morfemas de una lengua, junto con raíces complejas o palabras pre-formadas (o sea que no se arman en forma productiva), modismos y otras frases establecidas. Éstas son las estructuras lingüísticas que un hablante sabe como unidades completas y que puede usar sin tener que determinar sus significados a base de sus partes integrantes. Un diccionario (que a veces también se llama léxico) es un libro que exhibe elementos del léxico de una lengua, especialmente palabras, con una indicación breve de sus significados y usos. Contrástese con gramática; sintaxis, morfología, fonología, semántica. [Inglés:

	lexicon]
Lingüística computacional	La lingüística computacional puede considerarse una disciplina de la lingüística aplicada y la inteligencia artificial. Tiene como objetivo la creación e implementación de programas computacionales que permitan la comunicación entre el hombre y la computadora, ya sea mediante texto o voz.
Objeto	La gramática tradicional proporcionó los términos transitividad y objeto (tema de la siguiente sección), por el momento consideramos solamente la definición en el Diccionario de la Real Academia de la lengua Española: los transitivos son los verbos cuya acción recae en la persona o cosa que es término o complemento de la acción. De lo cual se define el complemento directo (objeto) como el complemento en el cuál recae directamente la acción del verbo, y el complemento indirecto (objeto indirecto) como la persona, animal o cosa en quien recae indirectamente la acción del verbo.
Oración	[1]: frase verbal junto con las frases nominales o adverbiales u otras oraciones que dependan de ella. Las oraciones pueden ser dependientes o independientes. Por ejemplo, la oración independiente "Dice Juan que te buscaba" contiene la oración dependiente "que te buscaba". A veces se marcan las oraciones poniéndolas entre corchetes. El estudio de la estructura de las oraciones es uno de los temas centrales de la sintaxis. [Inglés: clause]
Palabra	Es una raíz, junto con los afijos que dependan de ella y posiblemente de otras raíces (en el caso de una raíz compuesta), que puede pronunciarse sola en el uso normal de una lengua, por ejemplo, como respuesta a una pregunta. Frecuentemente las palabras tienen rasgos fonológicos especiales. En el náhuatl de Tetelcingo, por ejemplo, las palabras normalmente se pueden reconocer por el penúltimo acento. Compárese con frase, morfema. [Inglés: word]
Paradigma	Lista de formas relacionadas de una palabra, especialmente si son relacionadas por flexión. Por ejemplo: "hablo, hablas, habla" es un paradigma de formas de tiempo presente singular del verbo "hablar" en el español; "hablar, hablante, hablado" es un paradigma de formas infinitiva del mismo verbo. La yuxtaposición de paradigmas paralelos en forma de un cuadro, puede facilitar la comparación y por lo tanto el análisis de las formas. [Inglés: paradigm]

Patrones de manejo	Una traducción más adecuada para este término sería Patrones de Rección, pero para evitar la confusión con la misma palabra empleada en la Teoría de la Rección y el Ligamento de N. Chomsky, elegimos manejo sintáctico.
Pivote	Pivote con el sentido de álgebra de matrices.
Procesamiento de Lenguaje Natural (PLN)	Disciplina tiene por objetivo habilitar a las computadoras para que entiendan el texto, procesándolo por su sentido.
Ranura	Utilizamos éste término para referirnos a un espacio a ser llenado (“ <i>espacio en blanco</i> ”) dentro de una oración o frase. Slot, en inglés.
Recuperación de información	La recuperación de información es la ciencia encargada de buscar información en archivos de diversos tipos, en meta-datos y en bases de datos textuales, de imágenes o de sonidos. La plataforma sobre la cual es posible realizar dichas búsquedas se extiende desde computadoras de escritorio, redes de computadoras privadas o públicas hasta intranets e internet
Relación de dependencia	Una relación de dependencia sintáctica es aquella en la que una pareja de vocablos mantiene una relación de dependencia tradicional especificada por el árbol de dependencias sintácticas. Más explícitamente, las flechas que salen de un vocablo hacia otros se consideran los modificadores sintácticos del primero
Relaciones gramaticales	Aunque en la literatura de constituyentes se conocen como funciones o relaciones gramaticales, nosotros los denominamos de aquí en adelante como objetos sintácticos. El término argumentos se refiere a los complementos.
Relación sintáctica Semántica	Véase <i>relación de dependencia</i> La Semántica es el estudio de los significados de las estructuras de las lenguas (morfemas, palabras, frases, oraciones y otras). Una diferencia o semejanza semántica es una diferencia o semejanza de significados. Compárese con gramática, sintaxis, morfología, fonología, léxico. [Inglés: semantic, semantic]
Sintaxis	Es el estudio de cómo las palabras se combinan para formar frases y oraciones, ya sean dependientes o independientes. Compárese con gramática; contrástese con morfología, fonología, semántica, léxico. [Inglés: syntax]
Topicalización	En la topicalización se mueve un constituyente al inicio de la oración para hacer énfasis. Por ejemplo: Tortas como ésta, mi mamá nunca comería, donde tortas

como ésta va al final usualmente: mi mamá nunca comería tortas como ésta.

Wh

En el desplazamiento wh, se mueve un término inglés que comienza con wh al inicio de la oración para formar una interrogación.

[2]: a veces se usa en contraste con el término cláusula para indicar una oración independiente, con las cláusulas dependientes que pueda tener, o una serie de dos o más oraciones independientes coordinadas con una conjunción como "y" u "o".
Compárese con enunciado. [inglés: sentence]

Referencias

Referencias

- [Abney, 91] Abney, S. P. *Parsing by chunks*. In R. C. Berwick, S. P. Abney, and C. Tenny (Eds.) *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer, Dordrecht, 257-278, 1991.
- [Aho *et al*, 86] Aho, A. V., R. Sethi and J. D. Ullman. *Compilers. Principles, Techniques and Tools*. Addison Wesley Publishing Company, 1986.
- [Allen, 95] Allen, J. F. *Natural Language Understanding*. Benjamin Cummings, 1995.
- [Allerton, 84] Allerton, D. J. (1984). "Three or four levels of co-occurrence relations." *Lingua*, 63, 17-40.
- [Amsler, 89] Amsler, B. (1989). "Research towards the development of a lexical knowledge base for natural language processing." In *Proceedings, 1989 SIGIR Conference*. Cambridge, MA.
- [Apresyan *et al*, 89] Yuri D. Apresyan, Igor Boguslavski, Leonid Iomdin, Alexandr Lazurski, Nikolaj Pertsov, Vladimir San-nikov, Leonid Tsinman. 1989. *Linguistic Support of the ETAP-2 System (in Russian)*. Moscow, Nauka.
- [Arjona-Iglesias, 91] Arjona-Iglesias, M. *Estudios sintácticos sobre el habla popular mexicana*. Universidad Nacional Autónoma de México, 1991.
- [Basili, 94] Basili, R., *et al*. *A "Not-so-shallow" parser for Collocational Analysis*. In *Proceedings International Conference COLING-94*. August 5-9 Kyoto, Japan, 447-453, 1994.
- [Benson *et al*, 86] Benson, M., Benson, E., Ilson, R. *The BBI combinatory dictionary of English: a guide to word combinations*. John Benjamins Publishing Company, Philadelphia (1986).
- [Benson *et al*, 86b] Benson, M.; Benson, E.; and Ilson, R. (1986b). *The Lexicographic Description of English*. John Benjamins.
- [Benson, 89a] Benson, M. (1989a). "The collocational dictionary and the advanced learner." In *Learner's Dictionaries: State of the Art*, edited by M. Tickoo, 84--93. SEAMEO.

- [Benson, 89b] Benson, M. (1989b). "The structure of the collocational dictionary." *International Journal of Lexicography*, 2, 1-14.
- [Benson, 90] Benson, M. (1990). "Collocations and general-purpose dictionaries." *International Journal of Lexicography*, 3(1), 23-35
- [Berry-Rogghe, 73] Berry-Rogghe, Godelieve. *The computation of collocations and their relevance to lexical studies*. In Aitken, Adam J.; Bailey, Richard W., and Hamilton-Smith, Neil (Eds.) *The Computer and Literary Studies*. Edinburgh University Press, Edinburgh, United Kingdom, 103-112, 1973.
- [Bloomfield, 33] Bloomfield, Leonard. *Language*. Holt, New York, 1933.
- [Bolshakov & Gelbukh, 2001] Bolshakov, Igor, Gelbukh, Alexander. A Large Database of Collocations and Semantic References: Interlingual Applications. *International J. of Translation*, V.13, No. 1-2, 2001, pp. 167-187.
- [Bolshakov & Gelbukh, 2003] Igor A. Bolshakov, Alexander Gelbukh. 2003. On detection of Malapropisms by Multistage Collocation Testing. NLDB-2003, 8th Int. Conf. on Application on Natural Language to Information Systems. Bonner Köllen Verlag, pp. 28-41.
- [Bolshakov, 04] Bolshakov, Igor A. A Method of Linguistic Steganography Based on Collocational-Verify Synonymy. *Lecture Notes on Computing science*.
- [Bolshakov & Gelbukh, 2004] Bolshakov, Igor, Gelbukh, Alexander. *Computational Linguistics. Models, Resources, Applications*. Ciencia de la Computación. First Edition, México, 2004.
- [Borsley, 90] Borsley, R. D. *Welsh Passives*. In *Celtic Linguistics: Readings in the Brythonic Languages*, a Festschrift for T. Arwyn Watkins, ed. Martin J. Ball, James Fife, Erich Poppe, and Jenny Rowland. Philadelphia & Amsterdam: Benjamin. (Published as vol. 68 of *Current issues in Linguistic Theory*), 1990.
- [Brants, 2000] Brants, Thorsten. 2000. TNT—A Statistical Part-of-Speech Tagger. In: *Proc. ANLP-2000, 6th Applied NLP Conference*, Seattle.
- [Bresnan, 78] 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.
- [Bresnan, 82] Bresnan, J. W., editor. *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA. 1982.
- [Bresnan, 95] Bresnan, J. W. *Lexicality and Argument Structure 1*. Paris Syntax and Semantics Conference. October 12, 1995

- [Briscoe & Carroll, 93] Briscoe, E. and Carroll, J. *Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars*. Computational Linguistics, 19(1): 25--60, 1993.
- [Briscoe *et al.*, 2002] Briscoe, Ted. Jhon Carroll, Jonathan Graham and Ann Copestake. 2002. Relational Evaluation Schemes. In: Proc. Of the Beyond PARSEVAL Workshop, 3rd Internacional Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria, 4-8.
- [Calvo & Gelbukh, 2003] Hiram Calvo, Alexander Gelbukh. 2003. Natural Language Interface Framework for Spatial Object Composition Systems. *Procesamiento de Lenguaje Natural*, N 31; www.gelbukh.com/CV/Publications/2003/sepln03-2f.pdf.
- [Calvo & Gelbukh, 2004] Calvo Hiram, Alexander Gelbukh, Adam Kilgarriff. 2004. Acquiring Selectional Preferences from Untagged Text for Prepositional Phrase Attachment Disambiguation. In: Proc. NLDB-2004, Lecture Notes in Computer Sciences, N 3136, pp. 207-216.
- [Calvo *et al.*, 2005] Calvo Hiram, Alexander Gelbukh, Adam Kilgarriff. 2005. *Distributional Thesaurus versus WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment*. In *Computational Linguistic and Intelligent Text Processing(CICLing-2005)*. Lecture Notes in Computer Sciences N 3406, Springer-Verlag, pp. 177-188.
- [Calvo & Gelbukh, 2006] Calvo, Hiram and Gelbukh Alexander 2006. *DILUCT: An Open-Source Spanish Dependency Parser based on Rules, Heuristics, and Selectional Preferences*. UNISCON 2006.
- [Cano, 87] Cano Aguilar, R. *Estructuras sintácticas transitivas en el español actual*. Edit. Gredos. Madrid, 1987.
- [Cerdá, 75] Cerdá, Massó Ramón. *Lingüística Hoy. Colección "Hay que saber"*. 3ª Edición, Teide. Barcelona, España, 1975.
- [Chomsky, 57] Chomsky, N. *Syntactic Structures*. The Hague: Mouton & Co, 1957.
- [Chomsky, 65] Chomsky, N. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA. 1965.
- [Chomsky, 70] Chomsky, N. *Remarks on Nominalization*. In R. A. Jacobs and P. S. Rosenbaum (eds.), *Readings in English Transformational Grammar*. Waltham, Mass.: Ginn-Blaisdell, 1970.
- [Chomsky, 82] Chomsky, N. *Some Concepts and Consequences of the theory of Government and Binding*. MIT Press, 1982. Editada bajo el título de *La nueva sintaxis. Teoría de la rección y el ligamento*. Ediciones Paidós, 1988.

- [Chomsky, 86] Chomsky, N. *Knowledge of language: Its nature, origin and use*. Praeger, New York, 1986.
- [Choueka, 88] Choueka, Y. (1988). "Looking for needles in a haystack." In Proceedings, RIAO Conference on User-Oriented Context Based Text and Image Handling, 609-623. Cambridge, MA.
- [Church & Patil, 82] Church, K. and Patil, R. *Coping with syntactic ambiguity or how to put the block in the box on the table*. Computational Linguistics 8, 139-149, 1982.
- [Collins, 99] Collins, M. *Head-driven Statistical Models for Natural language parsing*. Ph.D. Thesis University of Pennsylvania.
- [Cowie, 81] Cowie, A. P. (1981). "The treatment of collocations and idioms in learner's dictionaries." *Applied Linguistics*, 2(3), 223--235
- [Cruse, 86] Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press. Cambridge, United Kingdom.
- [Dalrymple *et al*, 95] Dalrymple, M., Ronald Kaplan, J. T. Maxwell III and Annie Zaenen (eds.). *Formal Issues in Lexical Functional Grammar*. Stanford CSLI Publications, 1995.
- [DECIDE, 96] The DECIDE project. *Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora*. <http://engdep1.philo.ulg.ac.be/decide/> 1996
- [Dellenbaugh & Dellenbaugh 90] Dellenbaugh, D., and Dellenbaugh, B. (1990). *Small Boat Sailing, a Complete Guide*. Sports Illustrated Winner's Circle Books.
- [DG Website, 99] *DG Website Dependency-Based Approaches to Natural Language Syntax*. <http://ufal.mff.cuni.cz/dg/%20dgmain.html> 1999.
- [Earl, 73] Earl, Lois L. *Use of word government in resolving syntactic and semantic ambiguities*. *Information Storage and Retrieval*, 9, 639-664, 1973
- [Erbach & Uszkoreit, 90] G. Erbach and H. Uszkoreit. *Grammar Engineering: Problems and Prospects*. Report on the Saarbrücken Grammar Engineering Workshop. University of the Saarland and German Research Center for Artificial Intelligence. CLAUS Report No. 1, July 1990
- [Fillmore, 76] Fillmore, C. J. *Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences* 280, 20-32, 1976.
- [Fillmore, 77] Fillmore, C. J. *The case for case reopened in Syntax and Semantics*. In: Cole P., J. R. Harms (eds.). Vol. 8: *Grammatical Relations*. Academic Press, NY. 1977.

- [Franz, 96] Franz, A. *Automatic Ambiguity Resolution in Natural Language processing. An Empirical Approach*. Lecture Notes in Artificial Intelligence 1171. Springer Verlag Berlin Heidelberg, 1996
- [Fraser, 94] Fraser, N. *Dependency parsing*, PhD thesis, UCL, London, 1994.
- [Galicia-Haro *et al*, 99] Galicia-Haro Sofia N., Bolshakov I. A. y Gelbukh A. F. *Un modelo de descripción de la estructura de las valencias de verbos españoles para el análisis automático de textos*. 1999
- [Galicia-Haro, 2000] Galicia-Haro Sofia N., *Análisis sintáctico conducido por un diccionario de patrones de manejo sintáctico para lenguaje español*. Tesis doctoral, CIC, IPN, México, 2000.
- [Galicia-Haro *et al*, 2001] Galicia-Haro Sofia N., Gelbukh A. F. y Bolshakov I. A. *Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes*. J. Procesamiento de Lenguaje Natural, No 27, September 2001. SEPLN, Spain, 55-64, 2001.
- [Gazdar *et al*, 85] Gazdar, G., E. Klein, G. K. Pullum, and I. A. Sag. *Generalized Phrase Structure Grammar*. Oxford, Blackwell, 1985.
- [Gelbukh, 97] Gelbukh, Alexander. *Using a semantic network for lexical and syntactical disambiguation*. CIC-97, nuevas aplicaciones e Innovaciones Tecnológicas en Computación, Simposio Internacional de Computación, Mexico City, Mexico, pp. 352-366, 1997.
- [Gelbukh, 98] Gelbukh, A. *Lexical, syntactic, and referencial disambiguation using a semantic network dictionary*. Technical report. CIC, IPN, 1998.
- [Gelbukh, 99] Gelbukh, A. *Syntactic disambiguation with weighted extended subcategorization frames* Proc. PACLING-99, 1999, pp 244-249.
- [Gelbukh, 2000] Gelbukh, Alexander. *Computational Processing of Natural Language: Tasks, Problems and Solutions*. Congreso Internacional de Computación en México D.F., Nov 15-17, 2000.
- [Gelbukh & Sidorov, 2001] Gelbukh, Alexander y Sidorov Grigori. *La estructura de dependencias entre las palabras en un diccionario explicativo del español: resultados preliminares*, 2001
- [Gelbukh *et al*, 2003] Alexander Gelbukh, Grigori Sidorov, Francisco Velásquez. 2003. *Análisis Morfológico Automático del Español a Través de Generación*. Escritos, N 28, pp. 9 – 26.
- [Gladki, 85] A. V. Gladki. 1985. *Syntax Structures of Natural Language in Automated Dialogue Systems*(in Russian). No. 34, Spain.

- [Halliday, 66] Halliday, M. A. K. (1966). "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, 148-162. Longmans Linguistics Library.
- [Halliday, 67] Halliday, M. *Notes on transitivity and theme in English*. Journal of Linguistics 3, 37-82, 199-244, 1967.
- [Halliday, 68] Halliday, M. *Notes on transitivity and theme in English*. Journal of Linguistics 4, 179-216, 1968.
- [Hellwig, 80] Hellwig, P. *PLAIN - A Program System for Dependency Analysis and for Simulating Natural Language Inference*. In: Leonard Bolc, ed., *Representation and Processing of Natural Language*, 271-376. Munich, Vienna, London: Hanser & Macmillan, 1980.
- [Hellwig, 86] Hellwig, P. *Dependency Unification Grammar (DUG)*. In: Proceedings of the 11th International Conference on Computational Linguistics (COLING 86), 195-198. Bonn: Universit.,t Bonn, 1986.
- [Hindle, 93] Hindle, Donald y Rooth, Mats. *Structural Ambiguity and Lexical Relations*. Computational Linguistics, 19(1), 103-120, 1993.
- [Hirst, 87] Hirst, Graeme. *Semantic interpretation and the resolution of ambiguity*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, United Kingdom, 263. 1987.
- [Hudson 84] Hudson, R. A. *Word Grammar*. Oxford, Blackwell. 1984.
- [Hudson, 90] Hudson, R. A. *English Word Grammar*. Oxford: Blackwell. 1990
- [Hudson, 98] Hudson, R. A. (eds.) *Dependency and Valency*. An International Handbook of Contemporary Research. Berlin: Walter de Gruyter. <http://www.phon.ucl.ac.uk/home/dick/wg.htm> 1998.
- [Katz & Postal, 64] Katz, J. J. and P. M. Postal. *An Integrated Theory of Linguistic Descriptions*. Cambridge, Mass. MIT Press, 1964.
- [Kay, 73] Kay, M. *The MIND system*. In R. Rustin (ed.) *Natural language processing*. New York, Algorithmics Press, 155-188, 1973.
- [Ledo-Mezquita, 2005] Ledo-Mezquita, Yoel. *Recuperación de información con resolución de ambigüedad de sentidos de palabras para el español*. Tesis doctoral, CIC, IPN, México, 2005.
- [Lombardi & Lesmo, 98] Lombardi, V., L. Lesmo. *Formal Aspects and Parsing Issues of dependency theory*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, 787-793, 1998.

- [Lyons, 77] Lyons, J. *Semantics*. Cambridge, 1977.
- [McKeown & Radev, 98] Kathleen R. McKeown and Dragomir R. Radev. *Collocations*. Dale, R., Moisl, H., and Somers, H. (Editors), Handbook of Natural Language Processing. Marcel Dekker, 1998.
- [Meillet, 26] Meillet, Antoine. *Linguistique historique et linguistique générale*. Vol. 1. Champion, Paris, 351, 1926.
- [Mel'cuk & Zholkovsky, 70] Mel'cuk, I. A. and A. K. Zolkovsky. *Towards a functioning meaning-text model of language*. *Linguistics* 57: 10- 47, 1970.
- [Mel'cuk, 79] Mel'cuk, I. A. *Dependency Syntax*. In P. T. Roberge (ed.) *Studies in Dependency Syntax*. Ann Arbor: Karoma 23-90, 1979.
- [Mel'cuk, 81] Mel'cuk, I. A. 1981. Meaning-Text models: an recent trend in Sovietic linguistics. *Annual Review of Anthropology* 10, 27-62.
- [Mel'cuk, 88] Mel'cuk, I. *Dependency Syntax: Theory and Practice*. New York: State University of New York Press, 1988.
- [Montague, 70] Montague, R. *Universal Grammar*. *Theoria* 36: 373- 398, 1970.
- [Montague, 74] Montague, R. *Universal Grammar*. In Richard Thomason (eds.), *Formal Philosophy*. New Haven: Yale University Press, 1974.
- [Morales-Carrasco & Gelbukh, 2003] Morales-Carrasco, Raul and Gelbukh, Alexander. 2003. Evaluation of TnT Tagger for Spanish. Proc. 4th Mexican International Conference on Computer Science, México. IEEE Computer Society Press.
- [Nakhimovsky & Leed, 79] Nakhimovsky, A. D., and Leed, R. L. (1979). "Lexical functions and language learning." *Slavic and East European Journal*.
- [Navarro *et al.*, 2003] Navarro, Borja. Monserrat, Civit, M. Antonia Marti, R. Marcos, B. Fernández. Syntactic, Semantic and Programatic Annotation in Cast3LB. Shallow Processing of Large Corpora (SProLaC), a Workshop on Corpus linguistic, Lancaster, UK, 2003
- [Perlmutter, 83] Perlmutter, D. N. (ed.) *Studies in Relational Grammar I*. Chicago: University of Chicago Press, 1983.
- [Peters & Ritchie, 73] Peters, P. S. and R. W. Ritchie. *On the generative power of transformational grammars*. *Information Science*, 6, pp. 49 - 83, 1973.
- [Pollard, 84] Pollard, C. J. *Generalized Context-free Grammars, head Grammars and Natural Languages*. Ph. D. Thesis Departament of Linguistics. Stanford University, 1984.

- [Pollard & Sag, 87] Pollard, C. J. and I. A. Sag. *Information-based syntax and semantics*. CSLI Lecture notes series. Chicago University Press. Chicago II. Center for the Study of Language and Information; Lecture Notes Number 13, 1987.
- [Pollard & Sag, 94] Pollard, C. J. and I. A. Sag. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago & London, 1994.
- [Rambow & Joshi, 92] Rambow O., Joshi A. *A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena*. In: International Workshop on The Meaning-Text Theory, K. Henelt, L. Wanner (eds.) Arbeitspapier der GMD, No. 671, 1992.
- [Resnik & Hearst, 93] Resnik, P. and Hearst, M. *Syntactic ambiguity and conceptual relations*. In: K. Church(ed.) Proceedings of the ACL Workshop on Very Large Corpora, 58-64, 1993.
- [Sag & Wasow, 99] Sag, I. A. and Wasow, T. *Syntactic Theory: A Formal Introduction*. Center for the study of language and information, 1999.
- [Sag, et al, 2003] Sag, Ivan, Tom Wasow, and Emily M. Bunder. 2003. *Syntactic Theory. A Formal Introduction*(Second Edition). CSLI Publications, Standford, CA.
- [Salomaa, 71] Salomaa, A. *The generative power of transformational grammars of Ginsburg and Partee*. Information and Control, 18, pp. 227-232, 1971.
- [Seco, 72] Seco, M. *Gramática esencial del español*. Introducción al estudio de la lengua. Aguilar, 1972.
- [Sekine et al, 92] Sekine, S., Carroll, J. J., Ananiadou, S. and Tsujii, J. *Automatic Learning for Semantic Collocation*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 104-110, 1992.
- [Sells, 85] Sells, P. *Lectures on Contemporary Syntactic Theories*. CSLI Lecture Notes, Stanford, CA. Number 3, 1985.
- [Sharman, 89] Sharman, R. A. *An introduction to the Theory of Language Models*, IBM UKSC Report 204, 1989.
- [Smadja, 93] Smadja, F. A. *Retrieving Collocations from Text: Xtract*. Computational Linguistics 19.1: 143-176, 1993
- [Small, 87] Small, S. *A distributed word-based approach to parsing: Word Expert Parsing*. In Natural Language Parsing System. Edited by Bolc. Springer Verlag, 1987.
- [Sowa, 84] John F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Publishing Co. Reading, MA.

- [Steele, 90] Steele, J. *Meaning - Text Theory*. Linguistics, Lexicography, and Implications. James Steele, editor. University of Ottawa press, 1990.
- [Tapanainen *et al*, 97] Tapanainen, P., Järvinen, T., Heikkilä, J., Voutilainen, A. *Functional Dependency Grammar*. <http://www.ling.helsinki.fi/~tapanain/dg/> 1997.
- [Tesnière, 59] Tesniere, L. *Elements de syntaxe structural*. Paris: Klincksiek. (German: Tesniere, L. (1980): Grundzüge der strukturalen Syntax. Stuttgart: Klett-Cotta.) 1959.
- [Wilks, 75] Wilks, Yorick A. *Preference semantics*. In Keenan, E. L. III (Ed.), *Formal Semantics of Natural Language*. Cambridge University Press, 329-348, 1975.
- [Wilks *et al*, 93] Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. *Providing Machine Tractable dictionary tools*. In *Semantics and the lexicon* (Pustejowsky J. Ed.) 341-401, 1993.
- [Wilks, 98] Wilks, Yorick A. *Senses and texts*. In *Computers and the Humanities*, 1998.
- [Yarowsky, 93] Yarowsky, David. *One sense per collocation*. Proceeding of ARPA Human Language Technology Workshop, Princeton, New Jersey, 266-271, 1993.
- [Yngve, 55] Yngve, Victor H. *Syntax and the problem of multiple meaning*. In Locke, William N. and Booth, A. Donald (Eds.), *Machine translation of languages*. John Wiley & Sons, New York, 208-226, 1955.
- [Yuret, 98] Yuret, D. *Discovery of Linguistic Relations Using Lexical Attraction*. Ph. D. thesis. Massachusetts Institute of Technology, 1998.

Índice de términos

Índice de términos

Será agregado en la versión final.

Anexos

Anexos

Anexo 1: Muestras del corpus Cast3LB

El corpus Cast3LB es un archivo de texto plano que contiene aproximadamente 3,500 oraciones etiquetadas sintácticamente. Se presentan unos ejemplos.

```
(
(S
(S.F.C.co-CD
(S.F.C
(sn-SUJ
(espec.fp
(da0fp0 Las el))
(grup.nom.fp
(ncfp000 reservas reserva)
(sp
(preop
(sps00 de de))
(sn
(grup.nom.co
(grup.nom.ms
(ncms000 oro oro))
(coord
(cc y y))
(grup.nom.fp
(ncfp000 divisas divisa))))))
(sp
(preop
(sps00 de de))
(sn
(grup.nom
(np00000 Rusia Rusia))))))
(gv
(vmis3p0 subieron subir))
(sn-CC
(grup.nom
(Zm 800_millones_de_dólares 800_millones_de_dólares))))
(coord
(cc y y))
(S.F.C
(sn-CC
(espec.ms
(da0ms0 el el))
```

(grup.nom.ms
 (w 26_de_mayo 26_de_mayo)))
 (sn.e-SUJ *0*)
 (gv
 (vmii3p0 equivalían equivaler))
 (sp-CREG
 (prep
 (sps00 a a))
 (sn
 (grup.nom
 (Zm 19.100_millones_de_dólares 19.100_millones_de_dólares))))))
 (Fc , .))
 (gv
 (vmis3s0 informó informar))
 (sadv-CC
 (rg hoy hoy))
 (sn-SUJ
 (espec.ms
 (di0ms0 un uno))
 (grup.nom.ms
 (ncms000 comunicado comunicado)
 (sp
 (prep
 (spcms del del))
 (sn
 (grup.nom.ms
 (np00000 Banco_Central Banco_Central))))))
 (Fp . .))
 (
 (S
 (sn-SUJ
 (espec.mp
 (da0mp0 Los el))
 (grup.nom.mp
 (ncmp000 abogados abogado)
 (sp
 (prep
 (sps00 de de))
 (sn
 (grup.nom
 (np00000 Microsoft Microsoft))))))
 (gv
 (vmip3p0 aseguran asegurar))
 (S.F.C-CD
 (conj.subord
 (cs que que))
 (sn-SUJ
 (espec.fs
 (da0fs0 la el))
 (grup.nom.fs
 (ncfs000 división división)
 (sp
 (prep
 (sps00 de de))
 (sn

(espec.fs
 (da0fs0 la el))
 (grup.nom.fs
 (ncfs000 firma firma))))))
 (gv
 (vmic1s0 disminuiría disminuir))
 (sadv-CC
 (rg gravemente gravemente))
 (sn-CD
 (espec.fs
 (dp3cs0 su su))
 (grup.nom.fs
 (ncfs000 capacidad capacidad)
 (sp
 (prep
 (sps00 para para))
 (S.NF.C
 (infinitiu
 (vmn0000 competir competir))
 (sp-CC
 (prep
 (sps00 en en))
 (sn
 (espec.ms
 (da0ms0 el el))
 (grup.nom.ms
 (ncms000 mercado mercado)
 (sp
 (prep
 (sps00 de de))
 (sn
 (espec.mp
 (da0mp0 los el))
 (grup.nom.mp
 (ncmp000 programas programa)
 (s.a.mp
 (aq0mp0 informáticos informático))))))))))
 (Fp . .))

Las muestras presentadas corresponden a las siguientes oraciones:

“Las reservas de oro y divisas de Rusia subieron 800 millones de dólares y el 26 de mayo equivalían a 19.100 millones de dólares, informó hoy un comunicado del Banco Central”

“Los abogados de Microsoft aseguran que la división de la firma disminuiría gravemente su capacidad para competir en el mercado de los programas informáticos”

Anexo 2: Las reglas principales de la gramática extraída

Se extrajeron 2668 reglas del corpus Cast3Lb. A continuación se muestran las reglas extraídas que se repiten más de 10 veces (al lado izquierdo aparece la frecuencia de cada regla):

295	S ← S S	19	sn ← sn sn sn
126	grupnom ← pi sp	19	S* ← sp sn
106	sn ← sadv sn	19	S ← S S*
105	sadv ← rg sp	18	grupnom ← aq S
100	sadv ← sadv rg	18	INC ← S*
100	S ← conj sn	17	S* ← sn S
96	sp ← sadv sp	16	sa ← sadv sa
75	S ← S sn	16	grupnom ← Z sp
63	sn ← sn sn	16	S* ← sn sn
52	sn ← sn S	16	S* ← sadv sn
47	sa ← sa S	16	S ← S S S
42	grupnom ← pi S	15	espec ← dp di
40	S ← aq sp sp	15	espec ← di dp
39	S* ← sn sp	15	S ← conj sadv
39	S ← conj sp	14	sadv ← rg sadv
38	sp ← sp sp	14	S* ← S* S
34	sn ← sn sp	13	sadv ← rg sn
34	sadv ← sadv S	13	s ← s s
32	s ← s S	12	sn ← sn sadv
29	grupnom ← Zp sp	12	S ← aq sadv sp
29	espec ← di di	11	sp ← sp sp sp
29	S ← sp S	11	grupnom ← pi sp S
28	grupnom ← pn sp	11	grupnom ← Z sn
26	s ← sadv s	11	S* ← sadv S
26	S ← S sp	11	S ← sn S
25	S ← aq sn	11	S ← aq sadv
22	S ← sadv S	10	grupnom ← w S
21	grupnom ← pi s	10	espec ← di dd
20	sadv ← sadv rg sp	10	S* ← sn sadv
20	sadv ← rg S	10	S ← sn sp
20	S ← conj S	10	S ← conj S S

Banco_Central (del:@n Banco_Central){1}
 un (@di uno){1}
 informó (@vm informar){1}
 subieron (@vm subir){1}
 reservas (@n reserva){1}
 divisas (de:@n divisa){1}
 Rusia (de:@n Rusia){1}
 Las (@da el){1}
 800_millones_de_dólares (@Zm 800_millones_de_dólares){1}
 hoy (@rg hoy){1}
 comunicado (@n comunicado){1}
 Banco_Central (del:@n Banco_Central){1}
 un (@di uno){1}
 informó (@vm informar){1}
 equivalían (@vm equivaler){1}
 26_de_mayo (@w 26_de_mayo){1}
 el (@da el){1}
 19.100_millones_de_dólares (a:@Zm 19.100_millones_de_dólares){1}
 hoy (@rg hoy){1}
 comunicado (@n comunicado){1}
 Banco_Central (del:@n Banco_Central){1}
 un (@di uno){1}
 eran (@vs ser){65}
 informe (según:@n informe){65}
 el (@da el){65}
 19_de_mayo (@w 19_de_mayo){65}
 el (@da el){65}
 reservas (@n reserva){65}
 oro (de:@n oro){65}
 Banco_Central (del:@n Banco_Central){65}
 las (@da el){65}
 18.300_millones_de_dólares (de:@Zm 18.300_millones_de_dólares){65}
 eran (@vs ser){65}
 informe (según:@n informe){65}
 el (@da el){65}
 19_de_mayo (@w 19_de_mayo){65}
 el (@da el){65}
 divisas (@n divisa){65}
 Banco_Central (del:@n Banco_Central){65}
 las (@da el){65}
 18.300_millones_de_dólares (de:@Zm 18.300_millones_de_dólares){65}
 calculan (@vm calcular){114}
 activos (@n activo){114}
 divisas (en:@n divisa){114}
 poder (en:@n poder){114}
 Banco_Central (del:@n Banco_Central){114}

Los (@da el){114}
se (@p0 se){114}
dólares (en:@n dólar){114}
estadounidenses (@aq estadounidense){114}
calculan (@vm calcular){114}
activos (@n activo){114}
divisas (en:@n divisa){114}
poder (en:@n poder){114}
Ministerio_de_Finanzas (del:@n Ministerio_de_Finanzas){114}
el (@da el){114}
Los (@da el){114}
se (@p0 se){114}
dólares (en:@n dólar){114}
estadounidenses (@aq estadounidense){114}
depende (@vm depender){114}
valor (@n valor){114}
su (@dp su){114}
cambio (del:@n cambio){114}
oficial (@aq oficial){114}
rublo-dólar (@n rublo-dólar){114}
establece (@vm establecer){114}
que (@pr que){114}
Banco_Central (@n Banco_Central){114}
el (@da el){114}

Anexo 4: Muestras del diccionario extraído

Se extrajeron 46,698 colocaciones del corpus Cast3LB. A continuación se presentan algunas de las colocaciones extraídas con su frecuencia de lado izquierdo.

1 Ómnibus de clase	1 abrir COORD como
2 ácido ADJ nucléico	1 abrir COORD porque
1 ácido DET el	2 abrir COORD que
1 ácido DET su	1 abrir FECH uno_de_enero_de_1992
1 águila ADJ real	3 abrir PRON él
1 águila DET el	3 abrir PRON que
1 águila DET uno	1 abrir SUST Antonia
1 álbum DET el	1 abrir SUST CTU
1 álbum de Juegos_Florales	1 abrir SUST China
1 ámbito ADJ exterior	1 abrir SUST Esther_Tusquets
1 ámbito ADJ semántico	1 abrir SUST acceso
2 ámbito DET el	1 abrir SUST agujero
1 ámbito de religioso	1 abrir SUST banco
1 ángel ADJ exterminador	1 abrir SUST bar
2 ángel DET el	1 abrir SUST boca
1 ángel VERB incendiar	1 abrir SUST caja
1 ánimo DET su	1 abrir SUST cama
1 ánimo de matar	1 abrir SUST camino
1 árbitro ADJ cándido	1 abrir SUST campo
1 árbitro ADJ español	1 abrir SUST colecta
1 árbitro ADJ extranjero	1 abrir SUST compás
1 árbitro DET el	1 abrir SUST competencia
1 árbol ADJ frondoso	1 abrir SUST diario
1 árbol ADJ peinado	1 abrir SUST frasco
1 árbol DET el	1 abrir SUST fuego
1 área ADJ adversaria	3 abrir SUST mercado
1 área ADJ dicho	2 abrir SUST oficina
3 área DET el	1 abrir SUST ojo
1 área DET su	1 abrir SUST portón
1 área SUST córtex	1 abrir SUST taberna
1 área VERB procesar	1 abrir SUST tubo
1 área VERB ser	1 abrir SUST vacío
1 área del cerebro	1 abrir SUST variedad
1 abrir ADJ justo	1 abrir SUST ventana
1 abrir ADV Constantemente	1 abrir SUST vez
1 abrir ADV antes	1 abrir VERB convertir
1 abrir ADV de_par_en_par	1 abrir VERB empezar
1 abrir ADV hoy	1 abrir VERB formar
1 abrir ADV sólo	5 abrir VERB haber
1 abrir ADV también	1 abrir VERB incluir
1 abrir ADV un_poco	1 abrir VERB llegar

1 abrir VERB rebosar
 1 abrir VERB ser
 1 abrir VERB volver
 1 abrir a coquetería
 1 abrir a empresa
 3 abrir a operador
 1 abrir a país
 1 abrir a participación
 1 abrir a seducción
 1 abrir con El mismo_mar_de_todos_los_veranos
 1 abrir con ansiedad
 1 abrir con que
 1 abrir en
 1 abrir en Erfurt
 1 abrir en Este
 1 abrir en Plaza_Mayor
 1 abrir en capa
 1 abrir mediante diseminación
 1 abrir mediante lanzamiento
 1 abrir para apoyar
 1 abrir para dejar
 1 abrir por parte
 1 abrir sobre nación
 2 abrir sps00 a
 1 abroncar por faena
 1 abrumador ADV tanto
 1 abrumarse SUST lector
 1 absorber ADJ insertado
 1 absorber COORD cuando
 1 absorber SUST cantidad
 1 absorber VERB ser
 1 absorber a banesto
 1 absorber para impedir
 1 absorber por célula
 1 absorbido por intestino
 1 absorbido por tejido
 1 absorto ante despliegue
 1 abstinencia del poder
 1 abstracción ADJ posible
 1 absurdo ADJ coherente
 1 absurdo ADV completamente
 1 abuela DET su
 1 abuela DET tu
 1 abuela SUST Cayetana
 1 abuelo DET su
 1 abuelo de saber
 1 abundancia DET el
 1 abundante ADV ya
 1 abundar COORD como
 1 abundar DET el
 1 abundar NEG no

1 abundar SUST novela
 1 abundar VERB temer
 1 abundar en asunto
 1 abundar en campo
 1 abundar en trazo
 1 aburrido por falta
 1 aburrimiento DET el
 1 aburrimiento de0cn0 qué
 1 aburrir ADV profundamente
 1 aburrir PRON donde
 1 aburrir SUST superdotado
 1 aburrir VERB llegar
 1 aburrir sps00 a
 1 abusar COORD si
 1 abusar PRON se
 1 abusar de fuerza
 1 abusar de sátira
 1 abuso ADJ empresarial
 5 abuso DET el
 2 abuso de contratación
 1 abuso de lance
 1 abuso en utilización
 1 acabar ADJ convertido
 1 acabar ADV De_todas_formas
 1 acabar ADV a_la_postre
 1 acabar ADV ahí
 1 acabar ADV bien
 1 acabar ADV diagonalmente
 1 acabar ADV nunca
 1 acabar ADV siempre
 1 acabar ADV ya
 1 acabar COORD así_es_que
 1 acabar COORD cuando
 1 acabar COORD debidoáque
 3 acabar COORD que
 2 acabar NEG no
 11 acabar PRON él
 1 acabar PRON ese
 1 acabar PRON que
 3 acabar PRON todo
 2 acabar PRON uno
 2 acabar PRON yo
 1 acabar SUST Barcelona
 1 acabar SUST Darlin
 2 acabar SUST Nochebuena
 1 acabar SUST año
 1 acabar SUST día
 1 acabar SUST democracia
 1 acabar SUST discrepancia
 1 acabar SUST guerra
 1 acabar SUST libro

1 acabar SUST momia
 1 acabar SUST poder
 1 acabar SUST sueño
 1 acabar SUST tiempo
 1 acabar SUST tragedia
 1 acabar SUST vuelta
 2 acabar VERB acabar
 1 acabar VERB conformar
 1 acabar VERB cumplir
 1 acabar VERB dominar
 1 acabar VERB haber
 1 acabar VERB hacer
 1 acabar VERB penetrar
 1 acabar VERB señalar
 1 acabar VERB tomar
 1 acabar con Caronte_de_la_Estigia
 1 acabar con contrato
 1 acabar con detalle
 1 acabar con gol
 1 acabar con ilusión
 1 acabar con lista
 1 acabar con locura
 1 acabar con maleficio
 1 acabar con poder
 1 acabar con presa
 1 acabar con reserva
 1 acabar con trabajo
 1 acabar con unión
 1 acabar de detalle
 1 acabar de forma
 1 acabar de golpe
 1 acabar en
 1 acabar en año
 1 acabar en bar
 1 acabar en centro
 1 acabar en frigorífico
 1 acabar en laguna
 1 acabar en rincón
 1 acabar en santiamén
 1 acabar en_contra_de suponer
 1 acabar hasta gorro
 1 acabar hasta postergación
 1 acabar por quitar
 1 acaecer PRON que
 1 acaecer VERB deber
 1 acaecer en Sokombo
 1 acaecido DET el
 1 acaecido en 1971
 1 acaecido en 1972
 1 acaecido en Monte_Rainer
 1 acallar SUST voz

1 acariciar PRON él
 1 acariciar SUST espolón
 1 acarrear ADV constantemente
 1 acarrear SUST gemelo
 1 acarrear SUST saco
 1 acarrear VERB poder
 1 acarrear VERB ser
 1 acaso ADV también
 1 acaso SUST recuerdo
 1 acaso en sentido
 1 acatar NEG no
 1 acatar SUST amnistía
 1 acatar SUST juicio
 1 acatar SUST mando
 1 acatar a plenitud
 1 acceder al mercado
 1 acceder al tercero
 1 acceso ADJ directo
 1 acceso ADJ violento
 1 acceso DET cada
 1 acceso DET el
 1 acceso DET este
 1 acceso DET uno
 1 acceso SUST tributo
 1 acceso a ciudad
 1 acceso a información
 1 acceso a investigación
 1 acceso de tos
 1 accesorio ADV más
 1 accesorio DET uno
 1 accesorio de indumentaria
 1 acción ADJ culminado
 1 acción ADJ eficaz
 1 acción ADJ espléndido
 1 acción ADJ liberador
 2 acción ADJ libre
 1 acción ADJ ofensivo
 1 acción CIF 171
 1 acción DET cada
 1 acción DET dos
 8 acción DET el
 3 acción DET su
 1 acción DET uno
 1 acción VERB consistir
 1 acción VERB prever
 1 acción de Alcan
 3 acción de Algroup
 1 acción de contragolpe
 1 acción de golpista
 1 acción de inspector
 1 acción del Komerčni_Banka

1 acción del aire
 1 acción del hombre
 1 acción del oxígeno
 2 acción en circulación
 1 accidentado VERB levantar
 1 accidentado VERB morir
 1 accidente ADJ atroz
 2 accidente ADJ laboral
 1 accidente ADJ ocurrido
 1 accidente ADJ sufrido
 3 accidente DET el
 3 accidente DET uno
 1 accidente de auto
 1 accidente de automóvil
 1 accionista ADJ mayoritario
 1 accionista ADJ principal
 3 accionista DET el
 1 accionista DET su
 1 accionista de Algroup
 1 accionista de sociedad
 1 acechar COORD mientras
 1 acechar COORD que
 1 acechar SUST blasfemia
 1 acechar a víctima
 1 acechar al belén
 1 aceite ADJ desnaturalizado
 1 aceite ADJ ideal
 2 aceite DET el
 1 aceite de colza
 2 aceite de oliva
 1 aceituna DET el
 1 aceituna del aperitivo
 1 aceleración del cohete
 1 acelerar ADV adicionalmente
 1 acelerar SUST reloj
 1 acelerar SUST velocidad
 1 acelerar VERB desarrollar
 1 acelerar VERB poder
 1 acelerar a vez
 1 acentuar ADV De_pronto
 1 acentuar PRON él
 1 acentuar PRON se
 1 acentuar SUST claridad
 1 acentuar SUST rasgo
 1 acentuar en personaje
 1 acepción ADJ analítico
 1 acepción ADJ embrollado
 1 acepción ADJ superficial
 1 acepción DET el
 1 aceptable ADV bastante
 1 aceptar ADJ importante
 1 aceptar ADV de_buen_grado
 1 aceptar ADV fácilmente
 1 aceptar ADV sólo
 1 aceptar COORD como
 1 aceptar COORD pues
 3 aceptar COORD que
 4 aceptar NEG no
 2 aceptar PRON él
 2 aceptar PRON se
 1 aceptar SUST aplazamiento
 1 aceptar SUST comienzo
 1 aceptar SUST factor
 1 aceptar SUST gente
 1 aceptar SUST injerencia
 1 aceptar SUST investigador
 1 aceptar SUST joven
 1 aceptar SUST responsabilidad
 1 aceptar SUST riesgo
 1 aceptar SUST suerte
 1 aceptar SUST trabajo
 1 aceptar VERB acabar
 1 aceptar VERB acostumbrar
 1 aceptar VERB convertir
 1 aceptar VERB haber
 1 aceptar VERB ir
 2 aceptar VERB poder
 1 aceptar VERB ser
 1 aceptar a él
 1 aceptar en demás