Instituto Politécnico Nacional

Centro de Investigación en Computación

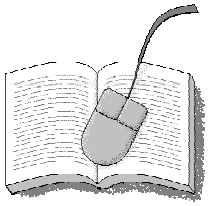
Análisis sintáctico conducido por un diccionario de patrones de manejo sintáctico para lenguaje español

TESIS

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

M. en C. Sofía Natalia Galicia Haro

Director de tesis: Dr. Alexander Gelbukh
Codirector: Dr. Igor Bolshakov



Laboratorio de Lenguaje Natural y Procesamiento de Texto

México, D. F. Año 2000

A G R A D E C I M I E N T O S

En una época donde el éxito de individuos de rasgos excepcionales o de excepciones individuales producidas por esfuerzos y perseverancias aisladas se presenta como ejemplo del éxito de algunos grupos humanos e incluso de las sociedades a las que pertenecen, donde decisiones individuales marcan la vida de ciudades e incluso de países, donde se cierran los caminos al análisis de decisiones que afectan a millones de personas, quiero patentizar que este trabajo es el resultado del acumulamiento de esfuerzos individuales pero principalmente de esfuerzos colectivos.

Agradezco la visión de los académicos que influyeron para que el IPN formara el CIC y en él tuviera cabida ampliamente el Laboratorio de Lenguaje Natural y Procesamiento de Texto, sin precedente en su tipo, y sobre todo, que invitaran al grupo de investigadores rusos que lo formaron y lo han seguido desarrollando. Su elevado nivel académico y su tradición gramatical permiten que los estudiantes mexicanos podamos profundizar y contrastar las teorías hoy clásicas en el área.

Todo el grupo del Laboratorio, profesores y estudiantes, ha influido en muy diversos grados en este trabajo, la mayoría muy directamente. De manera específica, agradezco al M. C. Manuel Montes y Gómez por su compañerismo y participación en este trabajo, también agradezco aunque de forma demasiado breve, de otra forma tendría que utilizar muchas páginas para expresar toda su ayuda y enseñanzas, mi agradecimiento más profundo al Dr. Alexander Gelbukh y al Dr. Igor Bolshakov, su trabajo de dirección y de codirección no pudo ser más comprometido. Les agradezco imbuirme en este mundo de la lingüística computacional, sus enseñanzas oficiales y *extra-aulas*, y por la confrontación con su preparación y dedicación.

Como estudiante, agradezco a todos los empleados académicos y administrativos del CIC que me apoyaron desde mi ingreso. Agradezco también a las compañeras y compañeros del Sindicato Mexicano de Electricistas que con su visión, hace muchos años, estipularon en el contrato colectivo los *Permisos especiales para beca*, permiso con que fui distinguida durante tres años. Sin ellos y sin los actuales miembros del SME, Jaime Novoa entre ellos, que todavía luchan por mantener las prerrogativas educativas, no me hubiera sido posible ser estudiante de tiempo completo, tampoco sin el apoyo del subdirector académico del CIC, M. C. Enrique Cauich, hubiera podido satisfacer en tiempo y forma con toda la documentación requerida para ese permiso.

Mi preparación en esos tres años fue apoyada por distintas becas, el primer año las becas CIC y PIFI, después, durante año y medio la beca DEPI y los últimos seis meses la beca del proyecto CONACYT 26424-A, a cargo del Dr. Igor Bolshakov. También, gracias al convenio ALFA-Cordial signado entre el IPN y la Comunidad Europea, tuve dos estancias en la Universidad Politécnica de Cataluña, en el Departamento de Lenguajes y Sistemas Informáticos, con la tutoría del Dr. Horacio Rodríguez Hontoria quien nos proporcionó el corpus empleado en esta tesis.

Por último, agradezco infinitamente, a mis padres Virginia y Alberto, por su amor, su apoyo y su confianza, a mi familia {{Eugenia}, {Rosalba, Rodolfo, Arturo, Víctor }, {Emma, Francisco, Emilio, Francisco }, {Virginia, Rodrigo, Alberto}, {María del Carmen, Rodrigo, Iván} por las alegrías y responsabilidades compartidas, y a mi familia elegida {Francisco Javier, Francisco Javier} por su ilimitada paciencia, apoyo incondicional y enseñanza de una forma de vida con libertad, comprensión y amor.

Cuando es verdadera, cuando nace de la necesidad de decir, a la voz humana no hay quien la pare. Si le niegan la boca ella habla por las manos, o por los ojos, o por los poros, o por donde sea. Porque todos, toditos, tenemos algo que decir a los demás, alguna cosa que merece ser por los demás celebrada o perdonada.

Eduardo Galeano

ABSTRACT

Syntactic analysis of Spanish language has been following the same research path of syntactic analysis of English language. What this work intends is obtaining an adequate model for syntactic structure acquisition and structure disambiguation for Spanish language analyzing some of their diverse characteristics.

In this thesis we review several formalisms chosen from the two main perspectives developed for syntactic analysis of natural languages: constituent grammars and dependency grammars. We analyze their subcategorization description and its relation to semantic roles or actants. We investigate the appropriate description of syntactic structures for Spanish language, a language with relaxed word order constrains, wide prepositional phrase use, direct object differentiated by prepositional phrase realization, and duplication of syntactic valences among other characteristics. We argue for the specific description of each predicative word: verbs, adjectives and nouns, as defined in dependency grammars .

Nowadays it is not possible to reproduce the way human beings disambiguate word links in a phrase. We share the idea that human beings employ different knowledge. Our syntactic structure acquisition model considers three types of knowledge: lexical, semantic and phrase structure. For syntactic ambiguity resolution we propose the classification of the output of the syntactic structure acquisition system composed of a module set. Each module is built based on a different method that represents a specific knowledge. Each module gives a set of weighted variants. Those weights are based on the satisfied characteristics in each method. So, each module gives a quantitative measure of the probability of each syntactic structure in a dependency structure format. To disambiguate syntactic structures a voting module uses the weights assigned in each module, voting for the maximum added value of variants. The result is a classified list of the syntactic variants.

The system includes government patterns module, semantic proximity module and extended CFG module. The three methods require the compilation of dictionaries: the advanced government patterns dictionary, the semantic network and the extended Context Free Grammar (CFG) rules. The advanced government patterns refers to lexical knowledge, the description of the arguments for predicative words, similar to that in the government pattern dictionary of the Meaning ⇔ Text Theory, associated to semantic valences. We propose an updated description computer-adequate and enriched with statistics of syntactic realization, statistics of diverse realization for the same valence and statistics of valences compatibility.

The CFG rules refers to phrase structure knowledge based on constituents. We create an extended CFG for Spanish language (with gender and number concordance) and we implement a chart parser. We assume equally weighted variants for the CFG module. The semantic network refers to semantic knowledge. When several structures are quite possible or adjuncts attachment is ambiguous the semantic proximity, i.e., the concepts more close related to the words in the possible constituents, could help to disambiguate structure variants. The idea behind the semantic proximity is finding the shortest paths between constituents obtained from the CFG module. It employs a

semantic network to give a measure of “semantic nearness” between constituents. For this purpose we assign different weights to relations, hierarchy concepts links and implicit relations.

In our model the advanced government patterns dictionary is the most practical to solve most of the structure ambiguities. The dictionary reflects the properties of the language itself, giving the syntactic constructions for each predicative word, i.e. the entire subcategorization information for each specific word. For Spanish, there are no dictionaries with complete subcategorization information. There is some spread information considered by several authors. For a big dictionary we require thousands of entries but manual work implies labor intensive and so much time. We propose a statistical method to compile the syntactic information.

We propose a method to compile the frequency of combinations. These combinations correspond to the specific predicative words and the prepositions that introduce their valences. In our dictionary, the weight of a combination is defined as the quotient of the frequency of the combination in the correct variants of parsing, i.e., in the texts, and its frequency in the incorrect variants of syntactic structures produced by the specific analyzer.

The statistical model to obtain those frequencies is based on two sources: one generating the true structures and one generating noisy variants which represent the parser’s mistakes. Thus, some combinations can have a weight greater than 1, which means that this combination appears in correct variants more frequently than in incorrect ones. Others can have a weight inferior to 1, which means that they more frequently appear in false variants. Finally, some combinations may have a weight of 1, which means that this combination is useless for disambiguation, even if it is frequent in the texts.

The statistical weights give actually the possibility to change the whole point of view on the nature and use of the dictionary that is used for the purpose of disambiguation, giving the kinds of errors that an analyzer makes. We obtained those weights by an iterative process. The process begins with an empty dictionary. For each phrase all hypothesis about syntactic structure made by the parser have the same weights in the first iteration. Once the frequency of the combination in the correct variants in the texts and its frequency in the incorrect variants of syntactic structure are determined for each founded combination a new weight calculation for all the variants is made. These steps are repeated until the difference between weights obtained in the previous iteration and the actual iteration is not greater than the established threshold.

Since such a dictionary should contain statistical weights of the combinations for specific words, we employ the method on the texts of the LEXESP Spanish corpus. We test the obtained results of the syntactic information compiled for the advanced government patterns dictionary on a group of 100 sentence extracted from that corpus and parsed by the CFG module. The true structures for the input sentences probed to be classified in the 35% rank of our experiments.

CONTENIDO

VISTA GENERAL DE LA TESIS

ÍNDICE DETALLADO DEL CONTENIDO

LISTA DE FIGURAS

INTRODUCCIÓN

MOTIVACIÓN Y RELEVANCIA

Lo que diferencia a los seres humanos de las bestias es su posibilidad de acaudalar el conocimiento comunicándolo de una persona a otra, de una generación a otra, de las épocas antiguas a las épocas modernas y a las épocas futuras. Esta comunicación se efectúa en la forma de lenguaje natural, siendo el español uno de los lenguajes más hablados del mundo. No sólo nos comunicamos con él, sino que almacenamos nuestro tesoro más valioso –el conocimiento de la raza humana– en la forma de lenguaje natural. El manejo eficiente de este conocimiento es vital para la humanidad en la época de la información.

Desde las épocas más antiguas existen las ciencias que estudian el lenguaje humano. Éstas se puede clasificar en tres grandes ramas. Unas estudian el lenguaje en comparación con otros lenguajes, observando las diferencias y semejanzas entre estos. Por ejemplo, ¿qué diferencias hay entre el español y el portugués? ¿Por qué el francés se parece más a español que el japonés? ¿Cómo se dice *libro* en alemán? Este grupo de ciencias incluye a las que estudian las lenguas nativas, tales como yaqui o nahua, sus diferentes dialectos, las costumbres y la cultura de la gente que los habla. También estudian diferentes dialectos del mismo lenguaje, por ejemplo: ¿cuáles diferencias hay entre el español de México y el de Argentina?

Otras ciencias estudian el lenguaje en comparación con su propio estado en las épocas antiguas. Por ejemplo, ¿cómo fue la transición del latín al español? ¿En qué siglo el sonido *x* (*sh*) en español se transformó a *j*, el proceso que dejó su relicto en el modo en que escribimos el nombre de nuestra patria, Mé*x*ico? ¿Cómo se va a transformar el español en los próximos siglos?

Finalmente, otras ciencias lingüísticas se dedican al estudio del propio lenguaje, de sus prefijos, raíces, sufijos, oraciones, y el sentido de las palabras, oraciones y párrafos. Cuáles palabras se escriben con acento y cuáles sin acento. Cuáles oraciones están bien formadas y cuáles no están escritas en buen español. Cuál estilo es apropiado para un cuento para niños, cuál para un artículo de periódico y cuál para un informe técnico.

México cuenta con los centros de estudios que hacen investigación en estas ramas de la ciencia lingüística, principalmente El Colegio de México. En estos centros, se cultivan los

estudios humanitarios, en vinculación cercana con las tradiciones seculares de historia, filología, antropología, sociología. Los resultados de estas investigaciones son los libros que nos ayudan a aprender otros lenguajes, a adquirir un buen estilo y cultura de escribir, a entender mejor nuestra historia a través del desarrollo histórico del lenguaje.

Sin embargo, hace unos 50 años surgió una nueva ciencia cuyo fin no tiene analogías en la historia. Se construyó una máquina destinada a ayudarnos –e imitarnos– en lo más humano que tenemos –en pensar. Es la computadora. Y como pensar y hablar es casi lo mismo, surgió la tarea de modelar el funcionamiento del lenguaje. No sólo describir el lenguaje, como lo hacen las ciencias humanitarias, sino modelarlo, construirlo –construir un modelo de lenguaje, un autómata que hable y entienda.

La nueva ciencia técnica que combina el conocimiento sobre la computación –las computadoras– y el conocimiento matemáticamente preciso sobre la estructura del lenguaje humano, se llamó lingüística computacional. Esta ciencia se encarga de todos los aspectos de la interacción de las computadoras y el lenguaje humano. La tarea final de esta ciencia –como la piedra filosofal de la alquimia– es la construcción de una máquina que hable y entienda como nosotros lo hacemos.

Los resultados de la lingüística computacional son programas de software. La diferencia entre las tareas y los métodos de la lingüística humanística y de la lingüística computacional se puede comparar con la diferencia existente entre el trabajo de un ornitólogo y un constructor de aviones: mientras el primero estudia el color de las plumas de diferentes pájaros y sus distintas áreas de vida, la tarea del segundo es construir –con los métodos matemáticos y de ingeniería– un pájaro de metal que vuele y ayude a volar al hombre.

Falta mucha investigación todavía para lograr construir una máquina que hable como las personas. Muy precisas y numerosas son las reglas que describen el lenguaje para esta tarea, previendo y minúsculamente describiendo para la máquina los fenómenos que parecen “obvios” para un humano. Inventando y desarrollando los formalismos en que esta descripción se puede hacer explícita. Desarrollando los algoritmos y las estrategias del manejo, dentro de la computadora, de esta cantidad enorme de información sobre el lenguaje.

Pero para ser útil, una máquina no tiene que entender todo lo que lee. Puede entender algo. Si sólo entiende sobre qué tema habla un texto (aunque no entienda qué quiere decir), nos facilita la búsqueda de los documentos en Internet sobre los temas que nos interesan. Por ejemplo: ¿Cuáles artículos discuten los problemas de democracia? O bien, si la máquina entiende algunos comandos en voz alta, le podemos dar estos comandos: “abre el archivo *informe.doc* y envíalo a mi jefe”. Incluso podemos dar estos comandos por teléfono, y escuchar la respuesta de la máquina. O bien, si la máquina entiende algunos hechos que se mencionan en el documento, puede –leyendo millones de archivos automáticamente– recopilarlos en una base de datos. Finalmente, puede traducir un archivo de un lenguaje a otro.

En los últimos 50 años –la época de las computadoras– la ciencia de la lingüística computacional ha visto un gran avance. Los países más desarrollados del mundo invierten millones y millones de dólares en el desarrollo de las herramientas y recursos para el procesamiento automático de sus lenguajes, siendo éstos en primer lugar el inglés, el japonés y el alemán, entre otros. Desgraciadamente, muy pocos grupos trabajan sobre el español, quedando así descuidada nuestra querida lengua y perdiendo su posibilidad de competir con los lenguajes de otros países. La mayoría de esos pocos grupos trabaja en España, contadas personas trabajan en este campo de importancia vital en América Latina.

Con el fin de acabar con esta triste situación, se fundó en el año de 1996 el Laboratorio de Lenguaje Natural y Procesamiento de Texto en el Centro de Investigación en Computación del Instituto Politécnico Nacional de México. El objetivo de este laboratorio es desarrollar las técnicas, las herramientas y los recursos para el análisis automático del lenguaje español, y la modelación por computadora de los procesos de comunicación en español de los seres humanos. El firme soporte de las autoridades del Instituto y del país –el CONACyT entre otras– ayudó al rápido crecimiento de la producción científica del Laboratorio. En 4 años, se publicaron aproximadamente 100 obras científicas, de nivel internacional en su mayoría. Se desarrollaron más de 20 proyectos de diferente escala. Se fundó una serie de congresos internacionales sobre lingüística computacional de reconocido prestigio, con la participación de los mejores especialistas a nivel mundial. Y en menos de 4 años, se presenta la defensa de la primera tesis de doctorado del Laboratorio.

La ciencia de la lingüística computacional se divide en sus ramas propias. En México, un grupo fuerte, en la UNAM, trabaja en el reconocimiento y generación de voz – los sonidos del habla; también trabajan en diálogos multimodales. En la UAM hay algunos desarrollos sobre la traducción automática. El tercer grupo, en el INAOE, Puebla –con el cual colaboramos intensivamente– desarrolla métodos de recuperación de información. En nuestro laboratorio, –además de intensos desarrollos prácticos– en el momento presente concentramos nuestros esfuerzos, en primer lugar en lo que consideramos más importante para el desarrollo a largo plazo: el análisis automático profundo del sentido, comunicado por los textos, y el desarrollo de los algoritmos, métodos y recursos para esto. Aunque el objetivo del Laboratorio es más amplio, el camino hacia este objetivo es a través de ciertos pasos; ésta es la fase que estamos desarrollando actualmente.

El primer paso para el análisis profundo del texto es su análisis sintáctico, la determinación de la estructura de cada oración. Es una tarea muy difícil de realizar con una alta calidad, y poco avance hay en el mundo en esta tarea. El problema más difícil que se enfrenta en el análisis sintáctico es la ambigüedad: la computadora encuentra más de una interpretación de cada oración y tiene que elegir una, la correcta. Muchos científicos consideran la resolución de la ambigüedad la tarea más importante en el análisis de lenguaje. Nuestro trabajo –y específicamente la presente tesis– está, de esta manera, en el corazón de los esfuerzos para lograr los objetivos de nuestra ciencia: habilitar a las máquinas para procesar, entendiéndola, la información escrita en el lenguaje natural español.

El plan con el cual atacamos el problema, consiste de una red de proyectos. En el

marco de cada uno de los cuales se desarrolla un módulo del futuro sistema completo. Ya se desarrolló el analizador morfológico que proporciona la entrada al analizador sintáctico. Ya se desarrolló el analizador sintáctico que proporciona las hipótesis del análisis sintáctico. Están en desarrollo los módulos que usan estas hipótesis para las tareas de recuperación y estructuración de información. Una parte del sistema serán los módulos que elijan las mejores hipótesis, es decir, resuelvan la ambigüedad de varios tipos. La tarea de desambigüación es compleja e incluye varios módulos de votación –los módulos que, con métodos diferentes, evalúen las variantes; la votación entre estos módulos independientes decide cuál variante es la mejor. El objetivo de la presente tesis es el desarrollo de uno de estos módulos evaluadores –el que se basa en el método del uso de los patrones estadísticos de manejo sintáctico para el español.

ÁMBITO

Lenguaje natural y lingüística teórica

El lenguaje se considera como un mecanismo que nos permite hablar y entender. Los lenguajes naturales[1], es decir, el inglés, el francés, el español, etc. son una herramienta genuina para la comunicación entre los seres humanos, ya sea en forma oral o escrita.

Actualmente, el avance tecnológico en los medios de comunicación impresos y electrónicos nos permite obtener grandes volúmenes de información en forma escrita. La mayoría de esta información se presenta en forma de textos en lenguajes naturales. Toda esa información contenida en los textos es muy importante ya que permite analizar, comparar, entender el entorno en el que vive el ser humano.

Sin embargo, se presentan dificultades por la imposibilidad humana de manejar esa enorme cantidad de textos. Entre las herramientas que ayudan en las tareas diarias, la computadora es, hoy en día, una herramienta indispensable para el procesamiento de grandes volúmenes de datos. Pero todavía no se logra que una máquina al capturar una colección de textos los comprenda suficientemente bien; por ejemplo, para que pueda aconsejar qué hacer en determinado momento basándose en toda la información proporcionada, para que pueda responder a preguntas acerca de los temas contenidos en esa información pero no explícitamente descritos, o para que pueda elaborar un resumen de la información.

Para lograr esta enorme tarea de procesamiento de lenguaje natural por computadora, analizando oración por oración para obtener el sentido de los textos, es necesario conocer las reglas y los principios bajo los cuales funciona el lenguaje, a fin de reproducirlos y adecuarlos a la computadora, incluyendo posteriormente el procesamiento de lenguaje natural en el proceso general del conocimiento y el

[1] Es un término ya adoptado que el lenguaje humano se denomine natural para diferenciarlo de los lenguajes artificiales en el área de la computación.

razonamiento.

El estudio del lenguaje, está relacionado con diversas disciplinas. De entre ellas, la Lingüística General es el estudio teórico que se ocupa de los métodos de investigación y de las cuestiones comunes a las diversas lenguas. Esta disciplina a su vez comprende una multitud de aspectos (temporales, metodológicos, sociales, culturales, de aprendizaje, etc.). Los aspectos metodológicos y de aplicación brindan los principios y las reglas necesarios en el procesamiento de textos.

Los principios y las reglas de la lingüística general, aunados a los métodos de la computación forman la Lingüística Computacional. Esta es la área dentro de la cuál se han desarrollado y discutido muchos formalismos adecuados para la computadora a fin de reproducir el funcionamiento del lenguaje con la finalidad de extraer sentido a partir de textos y viceversa, transformando los conceptos de sentidos específicos a los correspondientes textos correctos.

El proceso que se realiza con las herramientas proporcionadas por la Lingüística Computacional para realizar las tareas necesarias para pasar del texto a la estructura conceptual, y de ésta a los textos, lo denominamos, de aquí en adelante, proceso lingüístico de textos.

Proceso lingüístico de textos

El proceso lingüístico considera análisis y síntesis de textos, es decir, comprensión y generación de oraciones en lenguaje natural. Tanto en la generación como en la comprensión se realizan diferentes transformaciones o cambios de una estructura a otra para llegar al objetivo correspondiente, obtener los conceptos del texto o crear textos, respectivamente.

La generación de texto dentro de este ámbito empieza con la conceptuación del mensaje que se transmitirá y con la definición del nivel de generalización o de detalle en que se realizará. A continuación se sigue con la planeación de las estructuras. Los problemas específicos para construir estas estructuras están relacionados con las elecciones para representar un sentido específico, y con las elecciones de las estructuras particulares que se enlazan a las palabras. Existen otros criterios que intervienen en la construcción de la estructura, que no se consideran en el nivel de oración sino en el nivel del discurso completo, como la coherencia, expuesta mediante enlaces entre oraciones.

La comprensión en el proceso lingüístico, más compleja que la generación, parte de la representación de la información textual, es decir, de la cadena de palabras, y la traduce a diversas estructuras lingüísticas en varias etapas.

Las transformaciones que se requieren en el análisis y la síntesis son tan complejas que se dividen, tanto en la teoría como en la aplicación, en etapas generales. Para que la computadora realice estas etapas se requieren métodos adecuados para la descripción y construcción de las estructuras correspondientes, es

decir, se requieren formalismos lingüísticos de representación y computacionales.

En la lingüística general se considera que tres niveles generales componen el procesamiento lingüístico: la morfología, la sintaxis y la semántica. En el procesamiento lingüístico de textos, entre estos niveles, se elaboran descripciones y transformaciones computacionales de estructuras, al menos en dos etapas, en la primera a una estructura sintáctica y en la segunda a la estructura conceptual. Estos niveles no están totalmente delimitados, investigadores diversos difieren un poco en los puntos de vista para esta delimitación pero las diferencias no son cruciales.

Cada uno de los niveles, tanto en la generación como en la comprensión, tiene sus propias reglas y requiere colecciones de datos (diccionarios) apropiadas, aunque ciertas tareas pueden compartir recursos en el análisis y en la síntesis de textos. De hecho, en la construcción de recursos para el procesamiento lingüístico de textos un concepto importante es compartir recursos, dados los grandes esfuerzos que normalmente se requieren para su compilación.

Nuestra investigación se centra en el análisis y en el nivel sintáctico. Por lo que los niveles morfológico y semántico se consideran como los niveles adyacentes, cada uno apoyado en sus propias características. La sintaxis tiene estrechas relaciones con ambos niveles. En el nivel morfológico, las características que están relacionadas con el nivel sintáctico son las categorías gramaticales (las partes del habla y sus subclases), y algunas características morfológicas.

Las partes del habla (*part of speech* en inglés, POS) son: sustantivo, verbo, artículo, etc. En el análisis se realiza un marcaje de POS cuando se asignan estas categorías gramaticales a cada palabra dada, es decir, cuando se indica la función de cada palabra en el contexto específico de la oración. Este marcaje se hace considerando características morfológicas y sintácticas del lenguaje.

Las características morfológicas relacionadas con la sintaxis son las combinaciones que pueden caracterizar paradigmas. Los paradigmas aquí se refieren a los grupos de palabras relacionadas por su semejanza de significantes (la mínima forma significativa en la palabra) o por alguna relación entre sus significados (idea contenida en el significante). Entre las características morfológicas que caracterizan paradigmas están las formas de conjugación de los verbos (*am<u>o</u>, am<u>as</u>, am<u>a</u>, am<u>an</u>*, etc.), las variantes que expresan género y número de sustantivos, etc. Por ejemplo, la palabra *com<u>en</u>*, donde la inflexión *<u>en</u>* describe tiempo presente, modo indicativo, tercera persona del plural. Estas características se utilizan para relacionar palabras, frases u oraciones entre sí, es decir, para la coordinación; por ejemplo, del verbo con el sujeto (*ell<u>os</u> com<u>en</u>*), del sustantivo con el adjetivo (*cas<u>a</u> roj<u>a</u>*), etc.

Otra característica morfológica con repercusiones sintácticas y semánticas es la relacionada a las formas homónimas. Existen diferentes palabras morfológicas, como *banco, bancos*, que son variantes de un mismo lexema (la parte constante de una palabra variable que expresa la idea principal contenida) y existen formas homónimas de un lexema, con diferente sentido, que conforman un vocablo común.

Estas formas homónimas se numeran para describir sus sentidos. De esta forma, por ejemplo, se tiene *banco*$_1$ y *banco*$_2$, mientras el primero se refiere al sentido relacionado a guardar algo (*banco de ojos, banco comercial*), el segundo se refiere al sentido de asiento para una sola persona.

Formas homónimas como: *querer*$_1$ tener el deseo de obtener algo, y *querer*$_2$ amar o estimar a alguien, se distinguen por sus construcciones sintácticas, como se verá más adelante.

Sintaxis

La tarea principal en este nivel es describir cómo las palabras de la oración se relacionan y cuál es la función que cada palabra realiza en esa oración, es decir, construir la estructura de la oración de un lenguaje.

Las normas o reglas para construir las oraciones se definen para los seres humanos en una forma prescriptiva, indicando las formas de las frases correctas y condenando las formas desviadas, es decir, indicando cuáles se prefieren en el lenguaje. En contraste, en el procesamiento lingüístico de textos, las reglas deben ser descriptivas, estableciendo métodos que definan las frases posibles e imposibles del lenguaje específico de que se trate.

Las frases posibles son secuencias gramaticales, es decir, que obedecen leyes gramaticales, sin conocimiento del mundo, y las no gramaticales deben postergarse a niveles que consideren la noción de contexto, en un sentido amplio, y el razonamiento. Establecer métodos que determinen únicamente las secuencias gramaticales en el procesamiento lingüístico de textos ha sido el objetivo de los formalismos gramaticales en la Lingüística Computacional. En ella se han considerado dos enfoques para describir formalmente la gramaticalidad de las oraciones: las dependencias y los constituyentes.

ENFOQUE DE CONSTITUYENTES

Los constituyentes y la suposición de la estructura de frase, sugerida por Leonard Bloomfield en 1933, es el enfoque donde las oraciones se analizan mediante un proceso de segmentación y clasificación. Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso para cada parte dividiéndola en subconstituyentes, y así sucesivamente hasta que las partes sean las partes de la palabra indivisibles dentro de la gramática (morfemas).

La suposición de frase y la noción de constituyente, se aplica de la siguiente forma. La frase *los niños pequeños estudian pocas horas* se divide en el grupo nominal *los niños pequeños* más el grupo verbal *estudian pocas horas,* este último a su vez, se divide en el verbo *estudian* más el grupo nominal *pocas horas* y así sucesivamente.

En la perspectiva de constituyentes, la línea más importante de trabajo es la desarrollada por el eminente matemático y lingüística Noam Chomsky, desde los años cincuenta. [Chomsky, 57] dice que lo que nosotros sabemos, cuando conocemos un lenguaje, es un conjunto de palabras y reglas con las cuáles generamos cadenas de esas palabras.

Bajo este enfoque, aunque existe un número finito de palabras en el lenguaje, es posible generar un número infinito de oraciones mediante esas reglas, que también se emplean para la comprensión del lenguaje. Como una subclase, muy importante, de las gramáticas formales, estas reglas definen gramáticas independientes del contexto (*Context Free Grammars* en inglés, CFG). Sin embargo, existen al menos dos cuestiones principales cuando se trata de la cobertura amplia de un lenguaje natural: el número de reglas y la definición concreta de ellas.

El número requerido de reglas para analizar las oraciones de un lenguaje natural no tiene límite predeterminado porque debe haber tantas reglas como sean requeridas para expresar todas las variantes posibles de las secuencias de palabras que los hablantes nativos pueden realizar. En cuanto a la definición, se generan mucho más secuencias de palabras de las que realmente quieren producirse. Por ejemplo, una regla para definir grupos nominales en el español es: un artículo indefinido, seguido de un sustantivo y a continuación un grupo preposicional. Sin embargo, esta regla define tanto *la plática sobre la libre empresa* como **la solidaridad sobre la libre empresa*[2] siendo ésta última una secuencia no gramatical.

En este enfoque, una información importante para el análisis sintáctico es la definida como subcategorización, referida a los complementos que una palabra rectora puede tener y la categoría gramatical de ellos. Los complementos, en la lingüística general, se definen como palabras, o grupos de elementos lingüísticos que funcionan como una unidad que completa el significado de uno o de varios componentes de la oración, e incluso de la oración entera. Esta información se ha agrupado en patrones que describen la composición de los complementos posibles para diferentes verbos, conocida como marcos de subcategorización.

Principalmente se considera que los verbos son las palabras del lenguaje que requieren estos marcos de subcategorización, los cuales pueden ser de diferentes tipos, simples como grupos nominales, o más complejos como por ejemplo, el verbo *dar* que subcategoriza un grupo nominal y un grupo preposicional, en ese orden, *Da un libro a María*. También se considera que la descripción de los complementos puede realizarse en términos sintácticos o en términos semánticos.

En términos sintácticos, se describen por su estructura y partes del habla. Por ejemplo: *en diez pesos* es un grupo preposicional compuesto de preposición, adjetivo numeral y sustantivo, *en una tienda* también es un grupo preposicional pero compuesto de una preposición, un artículo y un sustantivo. En este caso, como tanto

adjetivo numeral seguido de sustantivo y artículo seguido de sustantivo forman un grupo nominal, el mismo marco: preposición seguida de grupo nominal, describe ambos complementos.

La descripción en términos semánticos, por no estar considerada en una forma ligada a la descripción sintáctica, en este enfoque, se ha complementado con los papeles temáticos. Estos papeles temáticos tienen su antecedente en los *casos,* que son relaciones abstractas semánticas entre los verbos y sus argumentos, establecida en la Gramática de Casos [Fillmore, 77]. Intentan explicar las diferencias en las distintas estructuras para un verbo, por ejemplo: *Juan rompió la ventana con el martillo*, *El martillo rompió la ventana*, *La ventana se rompió*. Con los papeles temáticos se establece que *Juan, el martillo y la ventana*, hacen el papel de *agente,* y el *martillo* en la primera frase es una herramienta.

Las combinaciones de los distintos complementos en la oración presentan otra complejidad. Por ejemplo, en la frase *Compró el niño un libro en diez pesos en la tienda XX a un lado del metro Juárez a un vendedor alto de mal humor,* existen seis grupos preposicionales (*en la tienda, del metro Juárez,* etc.) introducidos con solo tres preposiciones, *a, en, de,* y aparecen dos grupos nominales (*el niño, un libro*). Las posibles combinaciones no son aleatorias pero estos complementos o grupos lingüísticos pueden ir enlazados en diferentes combinaciones, unidos al verbo o a algunos sustantivos de los diferentes grupos de la oración, por ejemplo: *Compró el niño, Compró un libro, Compró en diez pesos, Compró en la tienda XX, Compró a un vendedor alto, la tienda XX a un lado del metro Juárez.*

Mientras para un hablante nativo es obvio cómo se relacionan los complementos, para una computadora son posibles todas las variantes: *Compró a un lado, Compró del metro Juárez, Compró de mal humor, el niño en la tienda XX*, etc.

ENFOQUE DE DEPENDENCIAS

El primer intento real para construir una teoría que describiera las gramáticas de dependencias fue el trabajo de Lucien Tesnière en 1959. Las dependencias se establecen entre pares de palabras, donde una es principal o rectora y la otra está subordinada a (o dependiente de) la primera. Si cada palabra de la oración tiene una palabra propia rectora, la oración entera se ve como una estructura jerárquica de diferentes niveles, como un árbol de dependencias. La única palabra que no está subordinada a otra es la raíz del árbol.

Es importante notar que la motivación de muchas dependencias sintácticas es el sentido de las palabras. Por ejemplo en la frase *Los niños pequeños estudian pocas horas*, las palabras *pequeños* y *pocas* son modificadores de atributo de las palabras *niños* y *horas* respectivamente, y *niños* es el sujeto de *estudiar*. Un rasgo muy importante de las dependencias es que no son iguales: una sirve para modificar el

[2] El asterisco marca aquí y en adelante que la frase no es gramatical.

significado de la otra, así la secuencia *los niños pequeños* denota ciertos niños, y *estudian pocas horas* denota una clase de estudio.

En el enfoque de dependencias, la línea de trabajo más importante es la desarrollada por el investigador Igor Mel'cuk desde los años sesenta, la *Meaning* ⇔ *Text Theory* (MTT). Para [Mel'cuk, 79], en la sintaxis se describen los medios lingüísticos por los cuales se expresan todos los participantes que están implicados en el sentido mismo de los lexemas.

Bajo esta perspectiva, la descripción de conocimiento lingüístico es primordial. La descripción de los medios lingüísticos con los que se expresan los "objetos" del lexema se insertan junto con él en un diccionario, de esta forma se conoce de antemano cómo se relaciona el lexema con los distintos grupos de palabras en la oración. Por ejemplo, para el lexema *plática* aparecerá que utiliza la preposición *sobre* para introducir el tema, que *solidaridad* utiliza la preposición *con*, y que el verbo *dar* emplea un sustantivo para expresar el objeto donado y para introducir el receptor emplea la preposición *a*. Estas descripciones se denominan patrones de manejo.[3]

Una cuestión principal cuando se trata de la cobertura amplia de un lenguaje natural, empleando los patrones de manejo, se refiere al establecimiento de todo este conocimiento lingüístico que no se basa en lógica y que por lo tanto conlleva el enorme trabajo manual de la descripción de la colección completa de todos los posibles objetos de las palabras específicas (verbos, sustantivos o adjetivos). Por ejemplo, establecer la manera en que el lexema *comprar* expresa los participantes, en la acción de hacer que alguna cosa pase de una persona o entidad, a ser propiedad de otra persona o entidad, a cambio de una cantidad de dinero.

Con la sola descripción sintáctica de los complementos no hay una manera de establecer reglas para la computadora que definan las preposiciones específicas de cada verbo, por ejemplo la preposición *en* para el verbo *comprar* y no un grupo preposicional introducido por la preposición *sobre*. Y aún cuando se especificara particularmente para el verbo *comprar* que un complemento se introduce con la preposición *en*, se tiene que diferenciar entre grupos preposicionales como *en diez pesos* que expresa la cantidad de dinero y otros grupos preposicionales que expresan otros sentidos como *en una tienda*. Esta diferencia que implica un descriptor semántico está contemplada en la MTT.

En la MTT se relacionan los participantes semánticos con los complementos del verbo, es decir, la valencia semántica con la valencia sintáctica. Por ejemplo, la realización sintáctica *en diez pesos* se refiere a la cantidad de dinero por la cuál se compró algo si está relacionado con *comprar* o se trata de la cantidad en la cuál

[3] 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.

disminuye un precio si se trata de *reducir*, etc. En la MTT, la idea es establecer las valencias, es decir, los participantes referidos a la acción del verbo en cuestión, establecer quién realiza la acción, a quién está dirigida, qué se hace, etc. Por ejemplo, en la acción de *beber*, los participantes son quién bebe y qué bebe; en la acción *comprar* los participantes son: quién compra, qué compra, en cuanto lo compra, a quién se lo compra.

En este enfoque, también se considera necesario establecer la diferencia de los complementos seleccionados semánticamente, de los que expresan las circunstancias en las que se da la acción, que se denominan circunstanciales. Los complementos circunstanciales están relacionados al contexto local de la oración pero no expresan participantes en la acción del verbo, añaden información no relacionada directamente al sentido del lexema. Por ejemplo, en la frase, *compró contra su voluntad un traje nuevo*, el grupo preposicional *contra su voluntad* expresa un modificador a la acción *comprar*, pero no es un participante de la acción del verbo.

Peculiaridades sintácticas del español

Existen características dependientes del lenguaje que simplifican o vuelven más compleja la relación entre los grupos de palabras. Reconocer las combinaciones posibles de los verbos y sus complementos es menos complejo cuando en el lenguaje existen posiciones fijas de ocurrencia de ellos. Sin embargo esto varía, la estructura de la oración en diferentes lenguajes tiene diversos órdenes básicos y diferentes grados de libertad en el orden de palabras. Por ejemplo, el inglés y el español tienen un orden básico sujeto-verbo-complemento (SVC).

Esto no quiere decir que siempre se cumpla ese orden. Algunos lenguajes, como el inglés, tienen un orden más estricto, otros, como el español, tienen un grado de libertad mayor. Por ejemplo, la oración en español *Juan vino* a *mi casa* (SVC) se acepta sintácticamente en las siguientes variantes: *A mi casa vino Juan* (CVS), *Vino Juan a mi casa* (VSC), *A mi casa Juan vino* (CSV), *Juan a mi casa vino* (SCV), *Vino a mi casa Juan* (VCS), por lo que los participantes de las acciones pueden ocurrir en distintas posiciones respecto al verbo.

En español, al igual que en algunos otros lenguajes, el uso de las preposiciones es muy amplio. Este empleo, origina una gran cantidad de combinaciones de grupos preposicionales, pero también sirve para diferenciar, en muchos casos, la introducción de los participantes de una acción. Por ejemplo, en la frase *Compró el niño un libro en diez pesos*, los hablantes nativos reconocen que se utiliza la preposición *en* para introducir la expresión del precio del artículo comprado.

En español, el uso de preposiciones permite introducir sustantivos animados en el papel sintáctico de objeto directo, distinguir entre significados de verbos, distinguir participantes. Realmente, la preposición *a* entre otros usos, sirve para

---

Chomsky, elegimos *manejo sintáctico*.

diferenciar el significado del complemento directo de algunos verbos, por ejemplo, *querer algo* (tener el deseo de obtener algo) y *querer a alguien* (amar o estimar a alguien). Si este conocimiento se omite en el nivel sintáctico entonces el análisis en el nivel semántico se vuelve más complejo. Esta información también es útil en la generación de lenguaje natural porque dado el sentido que se quiere transmitir existe la posibilidad de seleccionar la estructura precisa para él.

Otra peculiaridad del español es la repetición restringida de valencias. Por ejemplo en la frase: *Arturo le dio la manzana a Víctor*, dónde *le* se emplea para establecer a quién le dieron la manzana y el grupo preposicional *a Víctor* también representa al mismo participante. Otro ejemplo es: *El disfraz de Arturo lo diseñó Víctor*, donde tanto *lo* como *el disfraz de Arturo* corresponden al objeto directo de *diseñar*. Esta repetición se da en forma de pronombres y sustantivos. Las implicaciones léxicas y sintácticas en cuanto a que algunos verbos presentan estas estructuras, a que se deben relacionar las dos expresiones de valencias sintácticas con la misma valencia semántica, y a posibles diferencias semánticas, competen al análisis sintáctico.

Ambigüedades en lenguaje natural

La ambigüedad, en el proceso lingüístico, se presenta cuando pueden admitirse distintas interpretaciones a partir de la representación o cuando existe confusión al tener diversas estructuras y no tener los elementos necesarios para eliminar las incorrectas. Para desambiguar, es decir, para seleccionar los significados o las estructuras, más adecuados, de un conjunto conocido de posibilidades, se requieren diversas estrategias de solución en cada caso.

Relacionada a la sintaxis, existe ambigüedad en el marcaje de partes del habla, esta ambigüedad se refiere a que una palabra puede tener varias categorías sintácticas, por ejemplo *ante* puede ser una preposición o un sustantivo, etc. Conocer la marca correcta para cada palabra de una oración ayudaría en la desambiguación sintáctica, sin embargo la desambiguación de este marcaje requiere a su vez cierta clase de análisis sintáctico.

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

También existe ambigüedad en los complementos circunstanciales. Por ejemplo, en la frase *Me gusta beber licores con mis amigos*, el grupo *con mis amigos* es un complemento de *beber* y *no de licores*. Mientras un hablante nativo no considerará la posibilidad del complemento *licores con mis amigos*, para la

computadora ambas posibilidades son reales.

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

La ambigüedad es el problema más importante en el procesamiento de textos en lenguaje natural, por lo que la resolución de ambigüedades es la tarea más importante a llevar a cabo y el punto central de esta investigación. Debido a que existe ambigüedad aún para los humanos, no es una tarea de la resolución de ambigüedades lograr una única asignación de estructuras en el análisis sintáctico de textos, sino eliminar la gran cantidad de variantes que normalmente se producen. Con los resultados de esta tesis, logramos promover las variantes con mayor posibilidad de ser las correctas hacia el grupo inicial en la clasificación de las variantes sintácticas generadas para cada oración.

OBJETIVO

Esta tesis propone un modelo para resolver el problema del análisis sintáctico relacionado a la gran cantidad de variantes generadas cuando se analizan textos sin restricciones. El modelo considera un algoritmo de desambiguación basado en tres diferentes fuentes de conocimiento del lenguaje, de las cuales la fuente principal dirige el análisis mediante conocimiento lingüístico. El algoritmo de desambiguación sintáctica restringe la gran cantidad de variantes que normalmente se generan, así que la base del análisis sintáctico pasa de la tarea infinita de definir una gramática de cobertura total para el lenguaje, la forma tradicional, a la tarea principal de buscar los objetos de cada palabra.

La primera fuente de conocimiento es lingüística y se describe en una colección de patrones de manejo sintáctico que reúnen información de cómo las palabras del español especifican léxicamente sus objetos, la segunda fuente se describe en una gramática extendida independiente del contexto para el español, y la tercera fuente se basa en proximidad semántica entre palabras.

Para lograr este objetivo, primero analizamos las características del español, principalmente las que difieren de los lenguajes cuyo orden de palabras es más estricto, para describirlas bajo un enfoque generalizado de descripción de valencias, con mayor énfasis en el formalismo de la MTT. Basándonos en este análisis proponemos una forma nueva de descripción de los Patrones de manejo, la denominamos Patrones de manejo avanzados, con información cualitativa para el análisis sintáctico. Debido al conocimiento lingüístico que se requiere en dichos patrones, proponemos un método semiautomático de adquisición de esa información, a partir de un corpus de textos. Por último, proponemos un algoritmo para reducir el número de variantes posibles de análisis, es decir, de desambiguación sintáctica.

Por lo que la investigación descrita en esta tesis incluye nuevas contribuciones en los aspectos explicados en las siguientes secciones.

Aplicación del modelo de dependencias al español

Los formalismos para análisis sintáctico basados en constituyentes han sido más apropiados para el inglés, principalmente por su orden de palabras más estricto. Debido al apoyo y a la cantidad de investigadores que trabajan en esta línea, se ha aplicado a muchos otros lenguajes, aún cuando no comparten la mayoría de las características del inglés.

Los modelos de dependencias que representan una continuación de las tradiciones europeas antiguas en lenguajes con un orden de palabras más libre, se han orientado más hacia un trabajo descriptivo, por lo que se han empleado muy restringidamente y en pocos lenguajes. De entre los modelos de dependencias la *Meaning ⇔ Text Theory*, que representa la tradición gramatical rusa, es la teoría más desarrollada, por su sistema formal que en alcance y contenido es comparable con la escuela generativa, de constituyentes.

Al español solamente se han aplicado formalismos basados en constituyentes. Una lista de los trabajos realizados basados en dependencias se encuentra en [DG Website, 99].

La aplicación de la MTT al español permite describir algunas características del español de una manera más natural y adecuada, como el orden más libre de palabras (comparado con el inglés), el uso de palabras específicas para introducir complementos seleccionados semánticamente y también para establecer la relación entre valencias sintácticas y semánticas.

Algoritmo de adquisición de patrones de manejo

La aplicación de la MTT se ha realizado en forma limitada porque la compilación de los recursos necesarios, diccionarios principalmente, requiere un esfuerzo enorme, por la necesidad de descripción del lenguaje en términos lingüísticos en todos los niveles. Para eliminar esta desventaja elaboramos un algoritmo que emplea métodos estadísticos y lingüísticos.

Los métodos puramente lingüísticos tienen el defecto de requerir por mucho tiempo la participación de recursos humanos calificados. Los métodos estadísticos, se han empleado con buenos resultados, en diferentes líneas de investigación. Una área importante de aplicación para los métodos estadísticos es la adquisición de información léxica. Los sistemas basados solamente en métodos estadísticos no han logrado el éxito total para resolver la mayoría de los problemas de procesamiento de lenguaje natural para los cuales fueron aplicados, sin embargo han sido muy útiles, y combinados con conocimiento lingüístico han demostrado cierta superioridad.

En esta investigación se combinan métodos lingüísticos que permiten extraer estructuras sintácticas, y métodos estadísticos para la selección de variantes de estructuras con la finalidad de obtener los complementos de palabras específicas (verbos, adjetivos y sustantivos).

Compilación del diccionario de patrones de manejo

La compilación de un diccionario de patrones de manejo avanzados para el español permite abarcar una cobertura amplia del lenguaje porque reúne conocimiento puramente lingüístico que no es posible reproducir mediante razonamiento ni mediante algoritmos. Se han compilado muy pocos diccionarios de este tipo, principalmente porque se han compilado manualmente y porque los diccionarios desarrollados incluyen el modelo completo de la MTT.

La compilación de los patrones mediante el algoritmo lingüístico estadístico desarrollado permite incluir información estadística adicional para eliminar cierta ambigüedad en el análisis sintáctico y para favorecer determinadas realizaciones que aparecen con mayor frecuencia en corpus de textos, lo cual no ha sido considerado en compilaciones de este tipo de diccionarios.

Este diccionario es un recurso para el procesamiento del español que servirá tanto para el análisis como para la síntesis en el nivel sintáctico.

Algoritmo de desambiguación sintáctica

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

Diversos formalismos se han desarrollado para tener una cobertura total en el análisis sintáctico de lenguajes naturales, sin embargo la principal dificultad que se ha presentado es reconocer las estructuras reales de entre una enorme cantidad de variantes generadas en dichos análisis.

Se han propuesto métodos que utilizan un solo modelado del lenguaje, por ejemplo con gramáticas independientes del contexto (CFG), con gramáticas de estructura de frase generalizada, con gramáticas de adjunción de árboles (TAG), etc. También se ha propuesto la combinación de formalismos con estadísticas, por ejemplo CFG con probabilidades, TAG con probabilidades, entre otros.

El algoritmo de desambiguación sintáctica que aquí presentamos se basa en la transformación a una forma compatible de las variantes sintácticas generadas mediante diversos modelos del lenguaje, en la evaluación cuantitativa de ellas y finalmente en una votación que clasifique las variantes para determinar las de mayor posibilidad de ser las correctas. Este algoritmo emplea como base principal el diccionario y los pesos de los patrones de manejo.

ORGANIZACIÓN DE LA TESIS

En el capítulo uno presentamos los antecedentes para el desarrollo de la investigación sobre análisis sintáctico, los formalismos gramaticales que se han desarrollado dentro de la lingüística computacional y las herramientas requeridas. A partir del capítulo dos presentamos nuestras aportaciones. En el capítulo dos desarrollamos la aplicación del modelo de dependencias al español, en el capítulo tres presentamos nuestro algoritmo de análisis y desambiguación sintáctica, y en el capítulo cuatro el algoritmo de adquisición del diccionario de patrones de manejo sintáctico.

En el capítulo uno, en la primera sección, revisamos las gramáticas generativas y las estructurales en su evolución histórica. Por una parte, la evolución de las teorías derivadas de los constituyentes para superar los problemas generados por las transformaciones y cómo se paliaron estos problemas mediante las restricciones. Por otra parte las teorías derivadas de las dependencias y los formalismos desarrollados. Por último, la tendencia lexicista como la convergencia de ambas descripciones.

Después presentamos la descripción de las estructuras sintácticas de los objetos de las palabras según cada uno de los formalismos representativos para comparar la información que cada uno propone y el nivel en el que sitúa su descripción. En la tercera sección del capítulo uno presentamos los métodos estadísticos para detectar regularidades en las secuencias de palabras en las oraciones, y en la última sección la noción de redes semánticas como descripción de conocimiento semántico.

En el capítulo dos presentamos la descripción detallada de las valencias, las complejidades que se presentan, las peculiaridades semánticas y sintácticas del español que se describen en los patrones de manejo y ejemplos de estos patrones para verbos, sustantivos y adjetivos. Describimos la información que proponemos para los nuevos patrones de manejo y la descripción de su notación formal. Presentamos también las diferencias entre la descripción de valencias en los enfoques considerados.

Presentamos primero la descripción del modelo general de análisis y desambiguación sintáctica, y posteriormente el algoritmo de compilación del diccionario ya que en ambos empleamos el analizador básico construido, basado en gramáticas generativas. Este analizador básico, representa una de las fuentes de conocimiento para el modelo general y en este contexto se describe detalladamente. En cambio, en la implantación del algoritmo de compilación del diccionario lo empleamos como herramienta de construcción de variantes.

En el capítulo tres describimos el modelo general de análisis sintáctico y desambiguación, propuesto, es decir, el modelo completo y cada uno de sus subsistemas. Describimos la gramática generativa experimental que desarrollamos, su creación, características y verificación. Presentamos el algoritmo seleccionado para realizar el análisis sintáctico con la gramática generativa. Describimos el algoritmo desarrollado para la transformación a una forma compatible de dependencias. Describimos también el empleo de la red semántica para la desambiguación sintáctica. Presentamos finalmente la formulación de la evaluación cuantitativa de las variantes sintácticas, el algoritmo de votación y su expansión a un multimodelo.

El algoritmo de adquisición de los patrones de manejo se describe en el capítulo cuatro. Presentamos primero la deducción del modelo, enseguida presentamos la evolución de su desarrollo, en su aplicación a textos modelados y posteriormente a textos reales, las estadísticas en ambos y su comparación. A continuación presentamos ejemplos de los patrones compilados, las estadísticas obtenidas y la comparación entre métodos de compilación en forma tradicional y en forma automatizada. Por último presentamos las pruebas realizadas sobre un conjunto de prueba para dar una medida de la efectividad del empleo del diccionario compilado.

Finalmente presentamos las conclusiones, que incluyen el motivo y las aportaciones de esta tesis, adicionalmente presentamos rumbos posteriores a esta investigación.

CAPÍTULO 1. RETROSPECTIVA HISTÓRICA DE LOS FORMALISMOS GRAMATICALES Y ALGUNAS HERRAMIENTAS EN LINGÜÍSTICA COMPUTACIONAL

1.1 GRAMÁTICAS GENERATIVAS Y LA TRADICIÓN ESTRUCTURALISTA EUROPEA

En muchas disciplinas, la retrospectiva histórica y el estado actual permiten una visión más clara de cada disciplina, desde el punto de vista de los principales enfoques y ejemplos representativos de cada una. Entonces presentamos de esta manera los formalismos gramaticales en la Lingüística Computacional. Consideramos los dos enfoques que por mucho tiempo se han considerado opuestos y que en años recientes tienen más coincidencias: la gramática generativa cuyo principal representante es la teoría desarrollada por Chomsky en sus diversas variantes, y la tradición estructuralista europea que proviene de Tesniére, con el ejemplo más representativo, la teoría Sentido ⇔ Texto de I. A. Mel'cuk. El sistema formal de esta última, en alcance y contenido es comparable con la escuela generativa.

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. Los enfoques que presentamos consideran esa jerarquía como relaciones entre combinaciones de las palabras o entre palabras mismas.

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 independientes del contexto, son gramáticas generativas también.

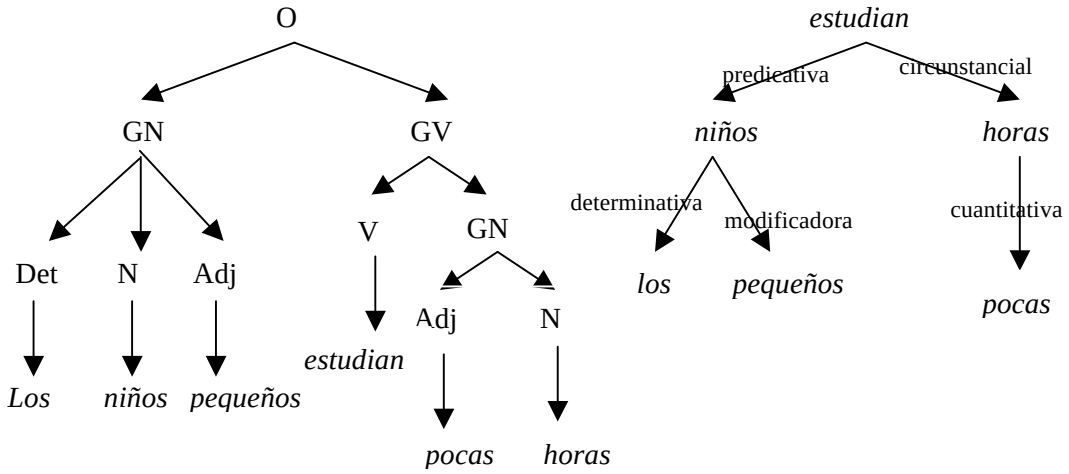
Chomsky y sus seguidores desarrollaron y formalizaron una teoría gramatical basada en la noción de *generación* [Chomsky, 65]. El trabajo que se realiza en la gramática generativa descansa en la suposición acerca de la estructura de la oración de que está organizada jerárquicamente en *frases* (y por consiguiente en estructura de frase). Un ejemplo de la segmentación y clasificación que se realiza en este enfoque se presenta en la Figura 1 A en el árbol de constituyentes para la frase *los niños pequeños estudian pocas horas,* donde O significa oración.

Un árbol de estructura de frase 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 estructura de frase, 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 GN (grupo nominal), GV (grupo verbal), N (sustantivo), GP (grupo preposicional), etc. aparecen en los árboles de estructura de frase 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 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.

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 actuantes) y auxiliares (denominados circunstanciales o modificadores). Las descripciones de valencias de



A) Árbol de constituyentes B) Árbol de dependencias

Figura 1. Estructuras sintácticas

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 *niños*, la modificadora es *pequeños*. 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 estructura de frase, emplean árboles a fin de describir la estructura de una frase u oración completa. Mientras la gramática de estructura de frase 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 1 B).

Las teorías de estructura de frase 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. Mientras los constituyentes han sido aplicados a la mayoría de todos los lenguajes naturales con la intención de una cobertura amplia, las dependencias han sido aplicadas en pocos lenguajes con una cobertura restringida. Primero presentamos un panorama del desarrollo de la estructura de frase y a continuación el desarrollo de las gramáticas con dependencias.

Gramática generativa en su primera etapa

VERSIÓN INICIAL
INCLUYENDO LA COMPONENTE TRANSFORMACIONAL

[Chomsky, 57], en su libro Estructuras Sintácticas, 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 1 A:

O → GN GV
GV → V GN
GN → Art Sust Adj
GN → Adj Sust

Adj → *pequeños* | *pocas*
Sust → *niños* | *horas*
V → *estudian*
Art → *los*

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 [57] 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. Las relaciones como sujeto y objeto[4], fueron un ejemplo de la necesidad del desarrollo de la gramática transformacional ya que su representación no era posible con las reglas independientes del contexto.

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. Ejemplos de transformaciones son el cambio de forma afirmativa a forma interrogativa, y de forma activa a pasiva.

La hipótesis de la gramática transformacional, es que por ejemplo, la frase (b) se deriva mediante reglas y el diccionario, de (a), con una transformación, alterando la estructura de tal forma, que la frase con *-qué* es inicial dentro de O.

(a) *Todos se preguntan* [*el profesor qué cosa ha dicho*]

(b) *Todos se preguntan* [*qué cosa ha dicho el profesor*]

Este tipo de transformación opera únicamente sobre oraciones que puedan analizarse con una estructura como

[4] 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.

$$\left[_{O}\ \text{X} - \textit{qué} - \text{NP} - \text{Y}_{v}\right]$$

donde *O* indica una oración, X una secuencia de palabras y Y_v el grupo verbal. GN es el grupo nominal.

En el ejemplo anterior *el profesor* correspondería a X y *ha dicho* correspondería a Y_v. La frase anterior entonces puede transformarse mediante la transformación que incluye el "movimiento" del constituyente X a la posición final, denotada como:

$$\left[_{O}\ \textit{qué} - \text{NP} - \text{Y}_{v} - \text{X}\right]$$

que corresponde a (b). Sin embargo, al estudiar más detenidamente el problema, encontramos que se requieren ciertas condiciones adicionales para la descripción general, por ejemplo el caso de X cuando es animado requiere la preposición *a*. Otra transformación es la que se realiza a partir de la estructura subyacente *el hombre está corriendo* para obtener la correspondiente forma interrogativa *¿Está corriendo el hombre*?.

Entre las transformaciones más importantes se encuentra la relacionada a las oraciones pasivas. Por ejemplo: *Un león fue atrapado por la policía*, que se deriva de las mismas estructuras subyacentes de sus contrapartes activas, *la policía atrapó un león*, por medio de una transformación a pasiva que permuta el orden de los dos grupos nominales e inserta las palabras *fue* y *por* en los lugares adecuados, directamente. En español, el cambio del objeto directo en la misma frase a una persona requiere además la inclusión de la preposición *a*, por ejemplo: *un ladrón fue atrapado por la policía* y *la policía atrapó a un ladrón.*

Otro punto muy importante de la GGT fue el tratamiento del sistema de verbos auxiliares del inglés, el análisis más importante en esta teoría. Chomsky propuso que el tiempo, en las formas verbales, estuviera en la estructura sintáctica subyacente, como un formante separado del verbo del cual formaba parte. Propuso dos transformaciones, una de movimiento para considerar la inversión del auxiliar en las preguntas y una de inserción que situaba *not* en el lugar apropiado para las oraciones de negación.

Ambas transformaciones, en algunos casos, tienen el efecto de un tiempo separado, es decir, lo dejan en una posición que no está adyacente al verbo. Para estos casos, Chomsky propuso una transformación para insertar el auxiliar *do* como un portador de tiempo. De esta misma forma se trataron, otros usos diversos del verbo auxiliar *do*, como la elipsis. Esta consideración unificada de aparentes usos diferentes de *do*, junto con la claridad formal de la presentación hicieron que muchos investigadores de la época se adhirieran a la GGT.

La GGT dominó el campo de la teoría sintáctica de los años sesenta a los

ochenta. La GGT cambió significativamente desde su aparición pero a pesar de su evolución, la noción de derivación transformacional ha estado presente de una u otra manera en prácticamente cada una de sus formulaciones.

TEORÍA ESTÁNDAR

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*, en inglés, 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 (tema de la sección 1.1.2). Otra aportación fue la adición de una componente semántica interpretativa a la teoría de la gramática transformacional.

Las reglas de estructura de frase permiten la recursividad, por ejemplo, en verbos como *decir* que además de tener un complemento tipo grupo nominal (*dijo una mentira*) aceptan complementos tipo oración (*dijo que María decía mentiras*). Un ejemplo de reglas recursivas es:

$$O \longrightarrow GN\ GV$$

$$GV \longrightarrow V\ O$$

En la primera regla, O puede reescribirse con GN GV, y a su vez GV tiene sustitución de O, y así sucesivamente (*Juan dijo que María dijo que Pedro dijo …*).

En la ST se presenta el concepto de *estructura profunda*, es decir, el árbol inicial en cada derivación de la oración. Esta estructura profunda representaba de una forma transparente toda la información necesaria para la interpretación semántica. Se sostenía que había un mapeo simple entre los roles semánticos desempeñados por los argumentos del verbo y las relaciones gramaticales [5] de la estructura profunda (sujeto, objeto, etc.). En el árbol final de la derivación, las palabras y las frases estaban ordenadas en la forma en que la oración sería realmente pronunciada, es decir, en su *estructura superficial*.

En esta teoría, las transformaciones se propusieron para ser el enlace primario entre voz y sentido, en el lenguaje. Los experimentos iniciales que mostraban una correlación entre la complejidad de una oración y el número de transformaciones propuestas en su derivación dieron credibilidad a esta idea pero investigaciones

[5] 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.

posteriores mostraron que no se podía sustentar. Ninguna teoría generativa actual mantiene esta idea central de las transformaciones.

Uno de los problemas fundamentales planteados por la ST es que el sentido está determinado a partir de la estructura profunda, antes de la aplicación de las transformaciones, pero entonces la influencia de las transformaciones sobre los sentidos no es nada clara.

La mayoría de las teorías gramaticales contemporáneas han mantenido las innovaciones más importantes de la ST, es decir, las características sintácticas, la estructura de frase recursiva y alguna clase de componente semántica.

TEORÍA ESTÁNDAR AMPLIADA

Chomsky y algunos otros abandonaron poco después de la ST la idea de que debían ser sinónimas las oraciones con estructuras profundas idénticas. En particular, demostraron que las transformaciones que reordenan grupos nominales cuantificados pueden cambiar el alcance de los cuantificadores. Un ejemplo muy conocido es el de *mucha gente lee pocos libros* que tiene interpretaciones diferentes de *pocos libros son leídos por mucha gente*. En consecuencia, propusieron que estructuras diferentes, de las estructuras profundas, debían desempeñar un papel en la interpretación semántica.

El marco teórico que Chomsky denominó Teoría Estándar Ampliada (*The Extended Standard Theory* en inglés, EST), propuso 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.

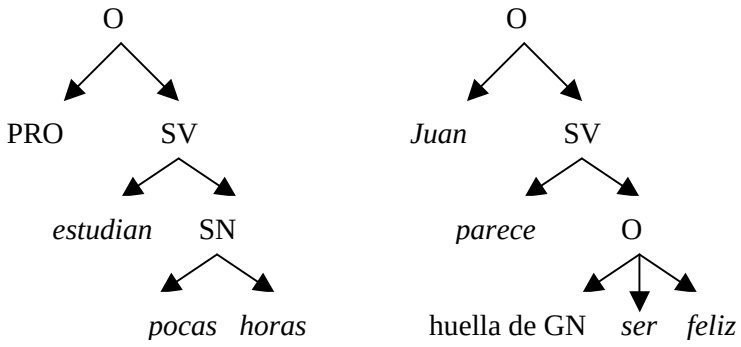


Figura 2. Categorías vacías

La gramática produce un conjunto de "pre-terminales" que no contienen más que marcadores gramaticales, marcadores de transformaciones (que indican cuales son las transformaciones que se efectuarán) y las categorías léxicas. Las reglas de inserción léxica reemplazan estas últimas por las palabras, produciendo así el conjunto de terminales.

La EST consideró la introducción de *categorías vacías*, que son elementos que ocupan posiciones en un árbol pero que no tienen una realización fonética. Incluyen un tipo de pronombre nulo usado en construcciones de control[6], y *huellas*[7] de elementos que han sido trasladados. Por ejemplo, ver Figura 2[8], un sujeto nulo (anáfora pronominal *pro*) en la frase española *Estudian pocas horas;* una huella de grupo nominal en la frase *Juan parece ser feliz* (la huella GN corresponde a *Juan*, el sujeto semántico de *ser*).

Uno de los intereses centrales de la EST y de trabajo posterior ha sido restringir la potencia de la teoría, es decir, restringir la clase de gramáticas que la teoría hace disponibles. La explicación principal para buscar esas restricciones ha sido considerar la posibilidad de la adquisición del lenguaje, la cuál fue considerada por Chomsky como la cuestión central de sus estudios lingüísticos.

[6] Construcciones para definir características de verbos como el inglés *try*. Por ejemplo en la frase *John tries to run (Juan intenta correr),* se considera que John es el sujeto de los verbos *try* y *run*.

[7] En inglés *trace*.

[8] Presentamos un árbol simplificado, omitiendo nodos intermedios de constituyentes.

Los sucesores y la paliación de los defectos del modelo transformacional

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, es decir, del tipo 0 en la jerarquía de Chomsky.

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.

[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.

En muchas versiones de la gramática transformacional, las oraciones pasivas y activas se derivaban de una estructura común subyacente, llevando a la sugerencia controversial, de que las derivaciones transformacionales preservaban muchos aspectos del sentido. 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.

TEORÍA DE LA RECCIÓN Y LIGAMENTO (GB)

La teoría de la Rección y Ligamento conocida como GB apareció por primera vez en el libro *Lectures on Government and Binding* de 81 [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 generación excesiva masiva que pudiera ocurrir.

La organización general de la GB con todos sus componentes[9], presentado por [Sells, 85] se muestra en la Figura 3.

Las estructuras -*d* y -*s* desempeñan una función similar pero no idéntica que las nociones de estructura profunda y superficial respectivamente de la ST. Estos niveles están relacionados por la operación Move-α, donde α se entiende que sea una variable sobre las categorías sintácticas. Puede considerarse que muchas de las transformaciones de las teorías precedentes se factorizaron en operaciones elementales donde ya no existen reglas específicas (transformaciones) como la de la pasiva sino que existe el movimiento de cualquier elemento a cualquier posición, y los principios y las restricciones regulan las operaciones de Move-α.

La Teoría θ (o de relaciones temáticas) provee información semántica. Los θ-roles se refieren a los participantes en la acción del verbo. En la GB se presupone que hay un número relativamente pequeño y por supuesto finito de estos roles, y emplea el criterio θ para establecer exactamente el número de argumentos que léxicamente especifica cada núcleo-*h*[10].

---

[9] Las subteorías y principios (proposiciones básicas o primarias) se marcan en rectángulos.

[10] De aquí en adelante núcleo-*h* 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

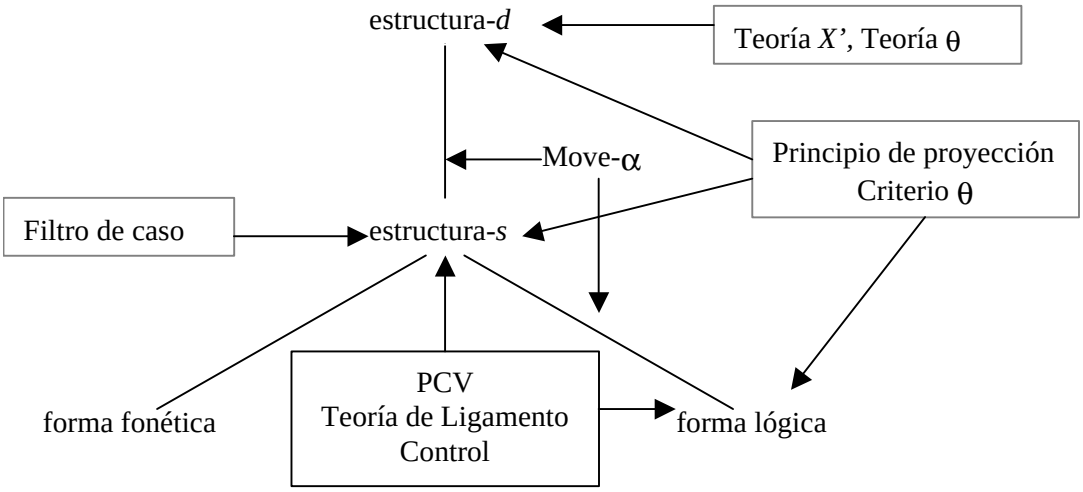


Figura 3. Organización de la GB

El filtro de caso se emplea para la buena formación de la estructura y la distribución de grupos nominales. Se basa en la tradicional noción de caso gramatical (nominativo, acusativo, dativo), que varía con el tipo de lenguaje.

La Teoría del Ligamento (*Binding Theory*, en inglés, BT) ha sido el mayor tópico de investigación dentro de la GB, caracteriza las relaciones interpretativas entre grupos nominales. La BT reúne principios como el Principio de la Categoría Vacía (PCV). El análisis en la GB propone diferentes tipos que podrían clasificarse de acuerdo a las características anafórica y pronominal, en abiertos o vacíos. Los de tipo abierto son explícitos y reflexivos; los vacíos son: desplazamiento *wh*[11] en formas interrogativas, pronombres tácitos del español (*pro*), pronombres para infinitivos (*PRO*), huellas de GN en verbos de control.

El movimiento va dejando huellas (una clase de categoría vacía), las cuales están limitadas por el elemento que se ha movido. La BT relaciona así las restricciones en el movimiento, con posibles relaciones de pronombres con antecedentes. La GB considera que, intuitivamente, las anáforas son aquellas que *deben* tener un antecedente (como los pronombres reflexivos) y los pronominales (como los pronombres personales) *pueden* tener un antecedente; todo esto se considera dentro de la misma cláusula. Puesto que el movimiento se usa para tratar con un rango amplio de fenómenos; entre ellos la relación activa - pasiva, la

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.

[11] En el desplazamiento *wh*, se mueve un término inglés que comienza con *wh* (*where*, *who*,etc.) al inicio de la oración para formar una interrogación.

extraposición[12], y la inversión de auxiliares, se produce un sistema abundantemente interconectado al ligar todos éstos a los principios de la BT.

En la GB hay un cambio importante en la descripción estructural. Las estructuras de frase están altamente articuladas, es decir, combinadas y relacionadas según ciertas normas de distribución, orden y dependencias. Distinciones y relaciones, lingüísticamente significantes, están codificadas dentro de las configuraciones del árbol tipo GB. Por ejemplo la categoría abstracta INFL, que contiene información de tiempo y concordancia, aparece en el árbol.

La literatura dentro de este formalismo es vasta, y representa un rango mucho más amplio de análisis que cualquiera de las otras teorías consideradas. Estudios lingüísticos del español se basan en este formalismo para sus descripciones [Lamiroy, 94], [Wilkins, 97].

El descendiente más reciente de la GB es el Programa Minimalista (PM) [Chomsky, 95]. Como su nombre lo implica, PM es más un programa de investigación que una teoría de sintaxis ya realizada. El PM explora la idea de que en lugar de generar oraciones directamente, lo que las gramáticas deberían hacer es seleccionar las mejores expresiones a partir de un conjunto de candidatas. El trabajo de elaborar los detalles del PM está aún en etapas iniciales.

Una diferencia conceptual mayor entre la GB y el PM es que en el PM los elementos léxicos portan sus características junto con ellos en lugar de asignárseles sus características basándose en los nodos en los que ellos rematan. Por ejemplo, los sustantivos llevan las características de caso con ellos y ese caso se revisa cuando los sustantivos están en una posición de especificación de concordancia.

El PM se origina a partir de la GB pero representa una considerable desviación del trabajo inicial en ese formalismo. Su meta es explicar la estructura lingüística en términos de condiciones de ahorro que son intuitivamente naturales en las gramáticas y en sus operaciones. Por ejemplo, los análisis tienen un mejor valor si minimizan la cantidad de estructura y la longitud de las derivaciones propuestas.

GRAMÁTICA DE ESTRUCTURA DE FRASE GENERALIZADA (GPSG)

La Gramática de Estructura de Frase Generalizada (*Generalized Phrase Structure Grammar*, en inglés, 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

[12] En la extraposición se mueven ciertos complementos tipo GN a la posición final de la oración, por ejemplo, la frase *Un niño brincando la cuerda fue visto* cambia a *Un niño fue visto brincando la cuerda*.

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* en inglés) que especifican solamente las frases que pueden aparecer como nodos en un árbol sintáctico, y las reglas de precedencia lineal (reglas LP, *Linear precedence* en inglés) 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. Por ejemplo:

O ⟶ GN GV

GV ⟶ duerme / Juan_

GN ⟶ Juan / _duerme

Las dos últimas reglas son reglas sensitivas al contexto, no generan nada porque la primera establece la reescritura de O por GN GV, pero ellas dos, interpretadas como la posibilidad de admisión, se refieren a que se admite *Juan duerme* como una oración a la que se le generaron árboles, enseguida se le revisaron los nodos y se verificó la cadena.

Así que aunque la GPSG excluye las transformaciones, la gramática se vuelve gramatical-léxica, pero realmente poco o nada se dice acerca del diccionario. Especialmente la información de subcategorías del verbo se encuentra en las reglas ID léxicas y no como entradas léxicas en el diccionario.

Esta teoría incluye la consideración del núcleo-*h* 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 la 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* en inglés) como: topicalización[13], 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. Por ejemplo, en la frase:

Which woman did Max say _ has declared herself President?

(¿Qué mujer dijo Max que se había declarado Presidenta?)

El guión bajo indica la posición de la frase desplazada *which woman*, que puede alejarse a una posición potencialmente sin límite en el árbol sintáctico. Mientras en la GB se dejaba una huella, en la GPSG el trato de este fenómeno involucra una codificación local de la ausencia del constituyente dado mediante una especificación de características.

Por ejemplo, a partir de la regla:

GV → H[40], O[FIN]

que introduce una oración finita como un nodo, se puede obtener, mediante una meta regla, la siguiente regla:

GV/GN → H[40], GV[FIN]

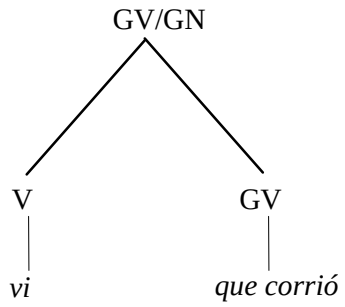


Figura 4. Fragmento de cláusula relativa

con un GV finito en lugar de la oración, y con la indicación del GN faltante mediante la diagonal. La GPSG incluye la introducción de *head* en las reglas, que se marca con

[13] 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*.

H en los ejemplos anteriores. La regla última permite el árbol sintáctico de la Figura 4, para un fragmento de la cláusula relativa *la niña que vi que corrió*, que correspondería al desplazamiento al inicio, de la cadena *la niña* en la frase *vi la niña que corrió*.

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.

GRAMÁTICA LÉXICA FUNCIONAL (LFG)

La teoría de la Gramática Léxica Funcional (*Lexical Functional Grammar* en inglés, 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 estructura de frase 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. Por ejemplo:

$$\begin{array}{lcccl} O & \longrightarrow & GN & GV & \} \text{estructura-c} \\ & & (\uparrow SUJ) = \downarrow & \uparrow = \downarrow & \} \text{estructura-f} \end{array}$$

Las flechas ($\uparrow$ y $\downarrow$) se refieren a la estructura-f correspondiente al nodo de la estructura-c construida por la regla. La flecha hacia arriba se refiere a la estructura-f del nodo madre y la flecha hacia abajo se refiere a la estructura-f del nodo mismo. Estas anotaciones indican que toda la información funcional que lleva el GN (es decir, la estructura-f de GN) va a la parte SUJ (sujeto) de la estructura-f del nodo madre (es

decir, la estructura-f de O), y que toda la información funcional que lleva el GV (es decir, la estructura-f de GV) también es información de la estructura-f del nodo madre. De esta manera se establecen las relaciones entre estructuras, la estructura-f para la frase *Paco come tacos*, sería la siguiente:

$$\begin{bmatrix} \text{SUJ} & [\text{PRED} \quad \textit{Paco}] \\ \text{OBJ} & [\text{PRED} \quad \textit{tacos}] \\ \text{TIEMPO} & \textit{PRES} \\ \text{PRED} & \textit{comer} < (\uparrow \text{SUJ})(\uparrow \text{OBJ}) > \end{bmatrix}$$

El valor de PRED (de predicado), indica el contenido semántico del elemento correspondiente. Por ejemplo el contenido semántico del sujeto en esa frase es *Paco*. En la entrada del verbo *comer* la parte léxica <(↑SUJ)(↑OBJ)> indica que el verbo subcategoriza un sujeto y un objeto; mediante las flechas se especifica que la estructura-f del nodo madre tiene un sujeto y un objeto. La inflexión del verbo añade la información del atributo tiempo verbal con el valor *PRES* (presente).

El nombre de la teoría enfatiza una diferencia importante entre la LFG y la tradición Chomskyana de la cuál se desarrolló: muchos fenómenos se analizan de una forma más natural en términos de objetos sintácticos (como se representan en el diccionario o en la estructura-f) que en el nivel de la estructura de frase. La parte léxica enfatiza la expresión para caracterizar procesos que alteran la relación de los predicados en el diccionario. Por ejemplo, la relación entre construcciones pasivas y activas.

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.

La regla de pasiva es una regla léxica, la cuál esencialmente añade el morfema de pasiva al verbo y cambia sus complementos de tal manera que el argumento asociado con el objeto de la forma activa se convierte en sujeto, y el sujeto se asigna a una función nula o a un Agente Oblicuo.

(SUJ) → ϕ / (OBL_{AG})

(OBJ) → (SUJ)

Por ejemplo, en la frase *tacos comidos por Paco*:

(↑PRED) = ‘comer< (↑SUJ) (↑OBJ) >’

| |

Agente Tema

| |

(↑PRED) = ‘comer< (↑OBL_{AG}) (↑SUJ) >’

En las LFG iniciales, la relación activa-pasiva fue codificada en términos de reglas léxicas, trabajo posterior 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.

GRAMÁTICA DE ESTRUCTURA DE FRASE DIRIGIDA POR EL NÚCLEO-H (HSPG)

La Gramática de Estructura de Frase dirigida por el núcleo-*h* (*Head-driven Phrase Structure Grammar* en inglés, 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 *hija* 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 núcleo-*h* 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 núcleos-*h* y algunas de las características de los nodos hijas 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 como la de la Figura 5 para la palabra *ella*, y enseguida mediante diagramas de matrices atributo-valor (MAV) en la **Figura** 6. En la Figura 5 las etiquetas de los nodos marcan los valores y las etiquetas de los arcos los atributos. En la **Figura** 6 los valores intermedios aparecen en la parte baja. Los cuadros marca $\boxed{1}$ establecen ligas de valores.

De acuerdo a principios especiales introducidos en la teoría, las características principales de los núcleos-*h* y algunas de las características de los nodos hijas 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-struc*) que representa la estructura de constituyentes inmediatos de la frase. El tipo *con-struc* tiene varios subtipos caracterizados por las clases de hijas que aparecen en las frases. El tipo más simple y más empleado es el *head-struc* que incluye HEAD-DAUGHTERS (HEAD- DTR) y COMPLEMENT-DAUGHTERS (COMP-DTRS), que a su vez tienen atributos PHON y SYNSEM. Por ejemplo para la frase *Eugenia corre* se tiene la estructura en la Figura 7.

Un punto importante en la HPSG es que tiene varios principios: de constituencia 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

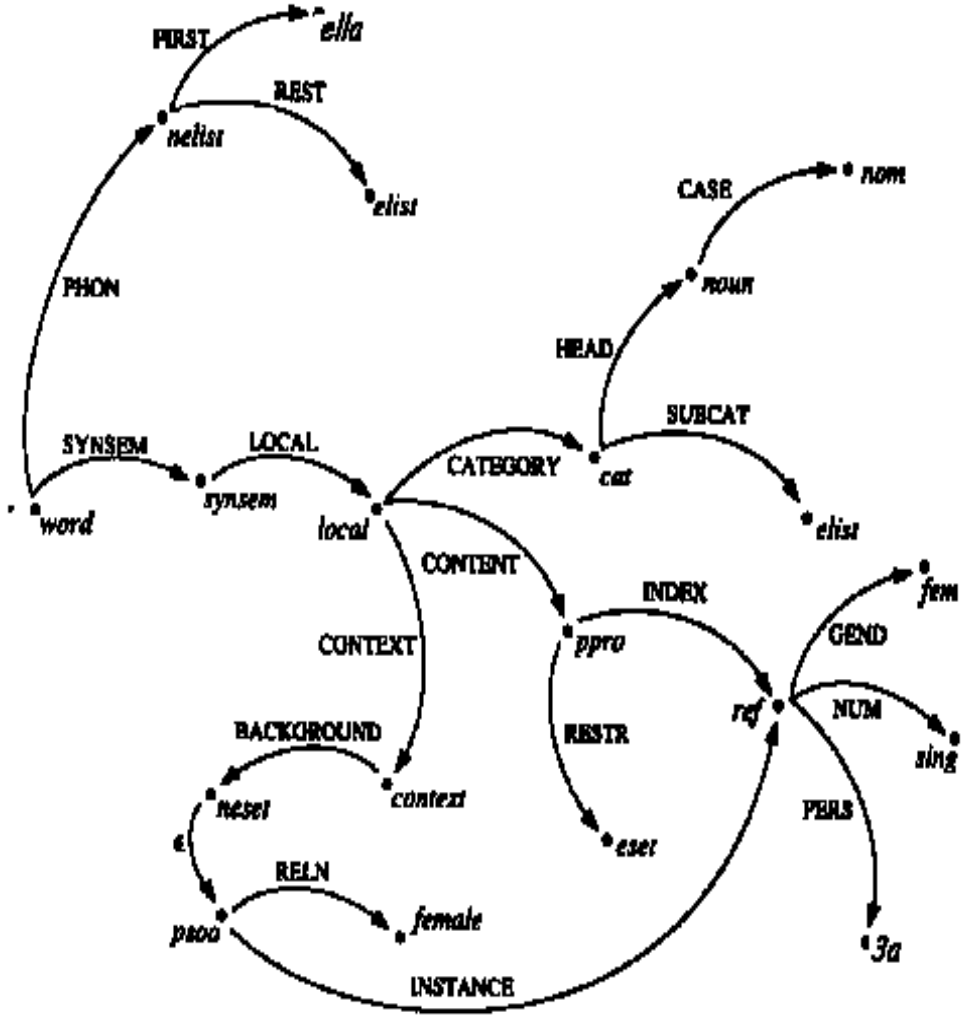


Figura 5. Estructura para el pronombre *ella*

particular dependen de las interacciones entre un sistema complejo de restricciones universales y particulares.

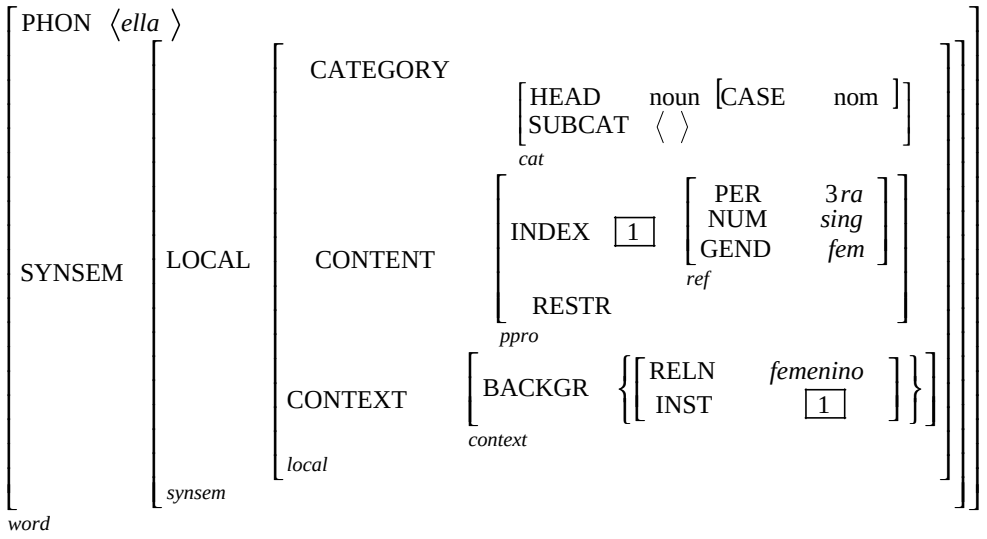


Figura **6**. Estructura de características mediante MAV

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

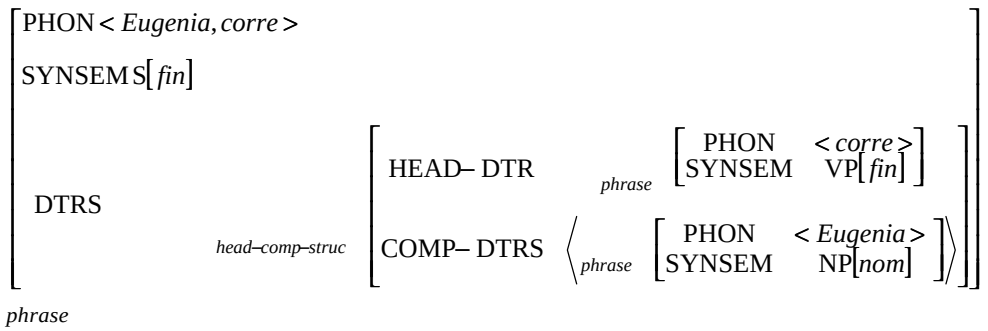


Figura 7. Estructura de características mediante MAV

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.

De las reglas a las restricciones

En contraste con la tradición de las gramáticas generativas hay otra aproximación a la teoría generativa, igualmente sometida a la meta original de desarrollo de gramáticas formuladas de manera precisa, las gramáticas basadas en la noción de satisfacción de restricciones en lugar de derivaciones transformacionales. En las gramáticas de restricciones las entradas léxicas incorporan información acerca de las propiedades de combinación de las palabras con la finalidad de que solamente se requieran operaciones generales esquemáticas en la sintaxis.

GRAMÁTICA CATEGORIAL (CG)

La Gramática Categorial (*Categorial Grammar*, en inglés, CG), introducida por [Ajdukiewicz, 35], adquirió importancia para los lingüistas cuando [Montague, 70] la usó como el marco sintáctico de su aproximación para analizar la semántica del lenguaje natural. La idea central de la CG es que una concepción enriquecida de categorías gramaticales puede eliminar la necesidad de muchas de las construcciones que se encuentran en otras teorías gramaticales (por ejemplo, de las transformaciones). Uno de los conceptos básicos de la CG, a partir de los setenta, es que la categoría asignada a una expresión debe expresar su funcionalidad semántica directamente, idea tomada de [Montague, 70].

Una gramática categorial consiste simplemente de un diccionario junto con unas cuantas reglas que describen cómo pueden combinarse las categorías [Wood, 93]. Las categorías gramaticales se definen en términos de sus miembros potenciales para combinarse con otros constituyentes, por lo que algunos autores ven a la CG como una variedad de la Gramática de Dependencias (tema de una sección posterior). Por ejemplo, las frases verbales y los verbos intransitivos pueden caracterizarse como aquellos elementos que cuando combinan con una frase nominal a su izquierda forman oraciones, una notación de esto es GN\O. Un verbo transitivo como *obtener* pertenece a la categoría de elementos que toman un GN en su lado derecho para formar una oración; esto puede escribirse (GN\O) /GN.

La suposición básica de la CG es que hay un conjunto fijo de categorías básicas, de las cuales se construyen otras categorías. Estas categorías básicas son:

sustantivo, grupo nominal y oración; cada una de las categorías básica tiene características morfosintácticas determinadas por el lenguaje específico. Para el inglés, el grupo nominal tiene características de persona, número y caso, el sustantivo sólo tiene número y la oración tiene forma verbal.

La CG no hace una distinción formal entre categorías léxicas y no léxicas, por lo que, por ejemplo, un verbo intransitivo como *dormir* se trata como perteneciente a la misma categoría que una frase consistiendo de un verbo transitivo más un objeto directo, como *obtiene un descanso*.

La operación fundamental [Carpenter, 95] es concatenar una expresión asignada a una categoría funcional, con una expresión de su categoría de argumento para formar una expresión de su categoría resultante; el orden de la concatenación está especificado como una categoría funcional. Por ejemplo, un determinante será especificado como una categoría funcional que toma un complemento nominal a su derecha para formar un resultante grupo nominal; la concordancia se maneja mediante la identidad de características simples.

La CG es esencialmente un formalismo de estructura de frase donde hay asignaciones léxicas a expresiones básicas y un conjunto de reglas de estructura de frase que combinan expresiones para producir frases totalmente basadas en categorización sintáctica. La CG difiere de otros formalismos en que postula un conjunto infinito de categorías y de reglas de estructura de frase en lugar de conjuntos finitos como en las CFG.

La atracción principal de la CG fue su simplicidad conceptual y por su adecuación a la formulación de análisis sintácticos y semánticos estrechamente ligados. Esto último debido a que se considera que restringe las asignaciones léxicas a expresiones básicas y a construcciones sintácticas potenciales, de tal forma que solamente se permiten las combinaciones de categorías sintácticas semánticamente significantes. Se asume en esta teoría, que la estructura sintáctica determina una semántica funcional manejada por los tipos de composiciones.

Se considera que por el empleo de las restricciones sintácticas y semánticas, todas las generalizaciones específicas del lenguaje se determinan léxicamente. Una vez definido el diccionario para un lenguaje, las reglas universales de combinación sintáctica y semántica se emplean para determinar el conjunto de expresiones gramaticales y sus sentidos. De lo anterior se observa la responsabilidad que se deja en el diccionario y que implica que deben proveerse mecanismos léxicos que consideren generalizaciones del lenguaje específico dentro del diccionario.

Una de las motivaciones para emplear este formalismo es la facilidad con que puede extenderse para proveer análisis semánticos adecuados de dependencias sin límite y construcciones de coordinación. La CG [Carpenter, 97] está grandemente influenciada por la LFG, la GPSG, la HPSG y otros análisis gramaticales categoriales y de unificación.

GRAMÁTICA DE RESTRICCIONES (GR)

En la Gramática de Restricciones (GR), Constraint Grammar en inglés [Karlsson *et al*, 95], toda la estructura relevante se asigna directamente de la morfología (considerada en el diccionario), y de mapeos simples de la morfología a la sintaxis (información de categorías morfológicas y orden de palabras, a etiquetas sintácticas). Las restricciones sirven para eliminar muchas alternativas posibles. Los autores indican que su meta principal es el análisis sintáctico orientado a la superficie y basado en morfología de textos sin restricciones. Se considera sintaxis superficial y no sintaxis profunda porque no se asigna ninguna estructura sintáctica que no esté en correspondencia directa con los componentes léxicos de las formas de palabra que están en la oración.

Ejemplos de esas restricciones para el inglés son:

- Una marca de verbo en presente, pasado, imperativo o subjuntivo, no debe ocurrir después de un artículo.
- La función sintáctica de un sustantivo en inglés es sujeto si va seguido de un verbo en forma activa y no intervienen sustantivos (de tipo sintáctico).

En la GR, la base de los postulados gramaticales son restricciones similares a reglas pero si el postulado gramatical falla se dispone de características probabilísticas opcionales. Para la GR tanto las restricciones (reglas gramaticales) como los postulados probabilísticos se requieren, no se trata de dos aproximaciones contrarias o de selección, aunque la relativa importancia probabilística es menor que en otras aproximaciones ya que aquí se enfatiza que el núcleo de la GR está destinado más a una naturaleza lingüística que a una probabilística.

Una idea relevante de la GR es poner en primer plano la descripción de ambigüedades, por lo que básicamente es un formalismo para escribir reglas de desambiguación. Divide el problema de análisis sintáctico en tres módulos: desambiguación morfológica, asignación de límites de cláusulas dentro de las oraciones y asignación de etiquetas sintácticas superficiales. Las etiquetas indican la función sintáctica superficial de cada palabra y las relaciones de dependencia básica dentro de la cláusula y la oración.

La noción de restricción se basa en hechos cercanos a la morfología superficial de la palabra, a la dependencia sintáctica entre palabras, y al orden de palabras, en lugar de basarse en principios abstractos de estructuramiento. La mayor desventaja es el trabajo necesario para establecer las restricciones, [Voutilainen, 95] postula 35 restricciones para desambiguar la palabra *that* y [Anttila, 95] emplea 30 restricciones sintácticas para la desambiguación del sujeto gramatical en inglés; los mismos autores postularon alrededor de 2000 restricciones para el inglés. La GR comparte con la LFG el uso de sujeto, objeto, etc. aunque como etiquetas que se toman del repertorio clásico de núcleo y modificadores, por lo que sus autores la

consideran *funcional*.

GRAMÁTICA DE ADJUNCIÓN DE ÁRBOLES (TAG)

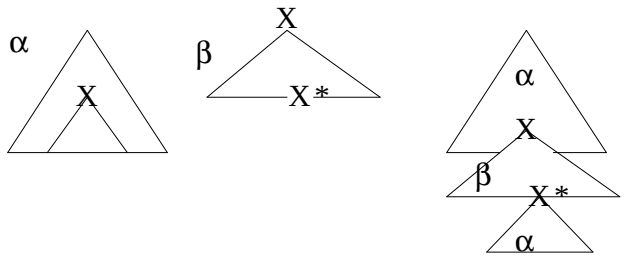
La Gramática de Adjunción de Árboles (*Tree Adjoining Grammar*, en inglés, TAG) [Joshi, 85] es una gramática definida por los elementos (I, A) donde I y A son conjuntos finitos de árboles elementales. Los árboles elementales están asociados con un elemento léxico, es decir, con una palabra, son una unidad sintáctica y semántica, y tienen operaciones de combinación. Estas operaciones tienen restricciones lingüísticas.

La TAG puede generar lenguajes más generales que las CFG pero no puede generar todos los lenguajes sensitivos al contexto, así que la fuerza de la TAG es ligeramente mayor que las CFG, en cuanto a las gramáticas que genera.

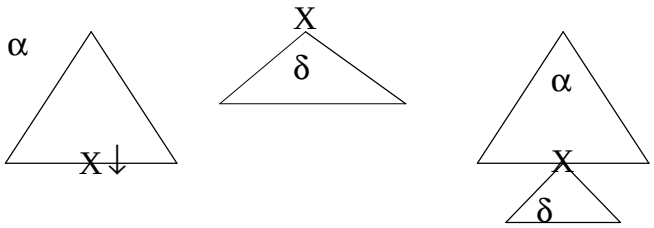
Los árboles iniciales tienen sólo terminales en sus hojas, y los árboles auxiliares se distinguen por tener un elemento X* en la base del árbol, cuya proyección es el nodo raíz X. La idea es que I y A sean mínimos en cierto sentido, que el inicial no tenga recursión en ningún no-terminal y que en los auxiliares, X corresponda a una estructura mínima recursiva que pueda llevar a la derivación si hay recursión en X.

Las operaciones son: adjunción y sustitución. La adjunción es una operación que separa un nodo interior del árbol inicial para adjuntar un árbol auxiliar. Al separar el nodo interior, el subárbol bajo éste, se transfiere a partir del elemento X*. La operación de sustitución simplemente sustituye un nodo hoja del árbol inicial por el árbol del auxiliar que se sustituye.

Operación de adjunción



Operación de sustitución



En la TAG, cada elemento léxico se llama ancla de la estructura correspondiente sobre la cuál especifica restricciones lingüísticas. Así que las restricciones son locales a la estructura anclada. Cada nodo interno de un árbol elemental se asocia con dos estructuras de rasgos: tope y bajo. La estructura-bajo contiene información relacionada al subárbol con raíz en el nodo (es decir, relación con sus descendientes), y la estructura-tope contiene información relacionada con al superárbol en ese nodo. Los nodos de sustitución tienen solamente una estructura-tope, mientras que los otros nodos tienen ambas estructuras: tope y bajo. En las dos operaciones definidas se unifican las estructuras de rasgos.

Gramáticas de dependencias.

[Mel'cuk, 79] explicó que un lenguaje de estructura de frase 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[14].

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* en inglés, 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

[14] No quiere decir que haya relaciones uno a uno del árbol de dependencias a las formas de palabras de la oración.

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 reacomodo estructural de grupos nominales.

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

SELECCIÓN SEMÁNTICA Y CONTEXTO LOCAL (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 Valencias 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 [Hellwig, 95] 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, ver como ejemplo [Briscoe & Carroll, 93].

La DUG ha sido implementada en el Instituto de Lingüística Computacional de la Universidad de Heidelberg como un marco de trabajo para análisis sintáctico de lenguajes naturales [Hellwig, 83]. Las DUG para el alemán, el francés y el inglés han sido elaboradas para los proyectos ESPRIT y LRE Translator's Work Bench (TWB) y Selecting Information from Text (SIFT).

Tres conceptos son los más importantes en esta teoría como gramática de dependencias: el lexicalismo, los complementos y las funciones. Por lexicalismo considera la suposición de que la mayoría de los fenómenos en un lenguaje dependen 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 el hecho de que cada complemento tiene una función

específica en la relación semántica establecida por el núcleo-*h*. 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 núcleo-*h* y un número de constituyentes que completan a ese elemento núcleo-*h*. 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.

DESCRIPCIÓN DEL CONJUNTO DE OBJETOS SINTÁCTICOS (WG, MTT)

Consideramos el conjunto de objetos sintácticos de los verbos como la variedad de marcos de subcategorización que pueden estar relacionados unos a otros a través de alternaciones de valencias. Pocos formalismos consideran todas las posibilidades de estas alternaciones como punto focal de su descripción sintáctica, entre ellos la Gramática de Palabra (*Word Grammar* en inglés, WG), y la Teoría Texto ⇔ Significado (*Meaning ⇔ Text Theory* en inglés, MTT).

GRAMÁTICA DE PALABRA

La Gramática de Palabra (en inglés, *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 sicología 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[15].

TEORÍA TEXTO ⇔ SIGNIFICADO

La Teoría Texto ⇔ Significado (en inglés, 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.

[15] Para el autor, la morfología se basa en estructura de constituyentes.

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 8 se muestran estos siete niveles como en [Mel'cuk, 88].

En la Figura 9, se presenta un ejemplo del árbol de dependencias de acuerdo a la MTT de [Mel'cuk, 88] para la frase *Siqueiros acusó a Rivera de pintar para turistas y esto agravó sus diferencias*, donde se hace una comparación con un árbol de constituyentes. Este árbol de dependencias presenta dos ventajas: requiere exactamente trece nodos (el número de palabras), el orden lineal de los nodos es

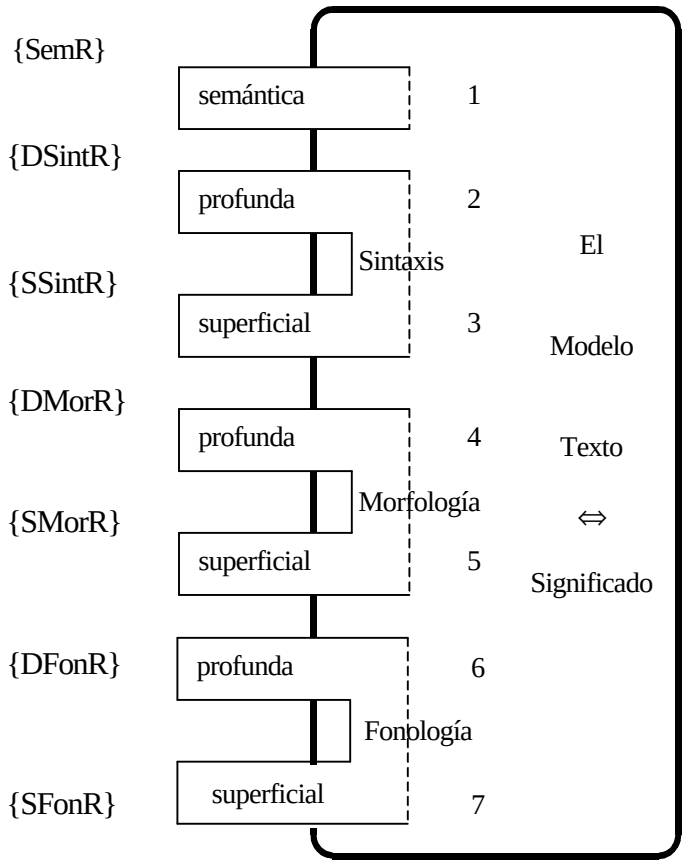


Figura 8. Niveles de Representación en la MTT

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 el léxico de un lenguaje dado. Los actuantes semánticos de una situación deben y

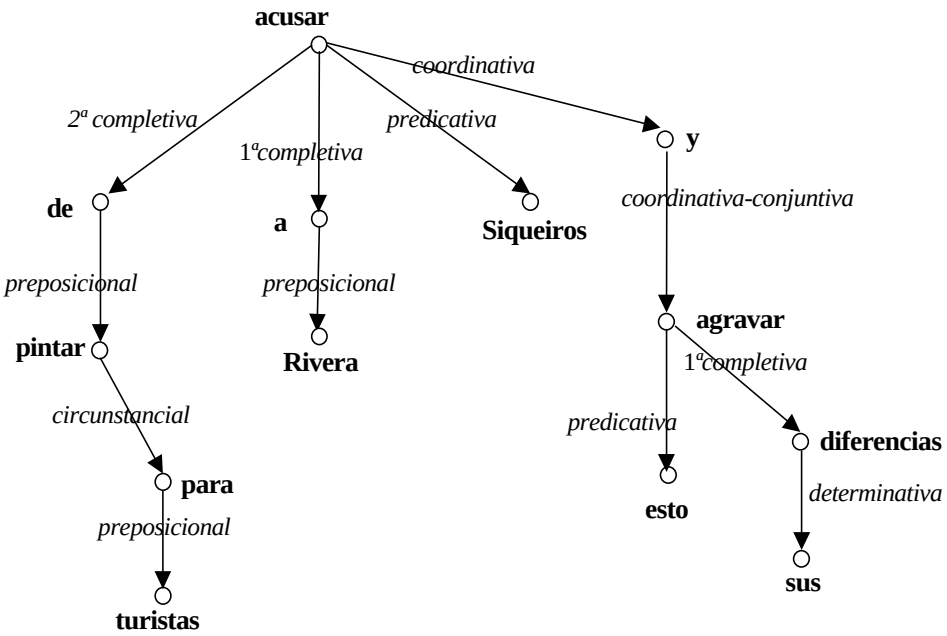


Figura 9. Ejemplo de estructura de dependencias en la MTT

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.

Métodos sin estructura sintáctica

Existen otros métodos en el procesamiento lingüístico de textos, considerados propiamente sin estructura sintáctica que están más orientados a los aspectos léxico y semántico, es decir, a los niveles morfológico y semántico. Si bien el interés de presentarlos aquí no se basa en información sintáctica, si se basa en presentar conceptos que dichos métodos aislaron de la sintaxis y que ahora se incorporan en los formalismos gramaticales más avanzados.

ANÁLISIS MEDIANTE EXPERTOS DE PALABRAS.

En [Small, 87], el autor presenta la teoría computacional Análisis Sintáctico mediante Experto de Palabra (*Word Expert Parsing* en inglés, WEP), una aproximación para entender el lenguaje natural como un proceso, de interacción de las palabras, distribuido y no uniforme.

La organización del WEP se basa en la creencia de que el agrupamiento de palabras para formar secuencias con sentido es un proceso activo que tiene éxito solamente gracias a la aplicación altamente idiosincrásica del conocimiento léxico, es decir, se fragmenta el texto y se comprende el significado de las piezas porque se conoce cómo las palabras particulares involucradas interactúan unas con otras.

Una excepción ocurre con las palabras que se ven por primera vez y entonces se aplica un proceso dirigido por la hipótesis, controlado por las palabras a su alrededor. Los autores insisten en el hecho de que las palabras son claramente la base de todos los fenómenos encontrados dentro del lenguaje.

Small basó su proposición en el hecho de que la estructura construida para la mayoría de los lenguajes naturales no había llevado de una forma directa a una interpretación semántica simple de las oraciones, basándose en su sintaxis.

La mayoría de las teorías de comprensión del lenguaje humano suponen que los lenguajes se basan en las regularidades (sintácticas, semánticas y conceptuales), en

cambio [Small, 87] y [Rieger & Small, 82] parten de una vista totalmente opuesta. Para ellos los sistemas de comprensión del lenguaje deben ser capaces de representar de forma más correcta las irregularidades que las regularidades.

En el WEP se considera a cada palabra como la fuente esencial del conocimiento necesario para la comprensión, de los conocimientos adecuados de sí misma y de sus relaciones con las otras palabras y conceptos. Para realizar esta tarea se liga un experto a cada palabra. El experto de cada palabra trata de determinar el rol significante de la palabra en el contexto, es decir, interactúa con otros expertos de palabras y con un modelo de proceso para adquirir el conocimiento conceptual apropiado para hacer las inferencias correctas.

Finalmente, los expertos para un fragmento de texto llegan a un acuerdo mutuo del significado del fragmento. Entre las funciones que realiza cada experto se encuentran las interacciones léxicas. En [Small, 87] los autores presentan un ejemplo de análisis, para la frase *The man throws in the towel* (el hombre echó dentro la toalla), ejemplo en el cual el experto más importante es el experto del verbo (*throws*) que construye un concepto de acción y refina su significado basándose en el contexto y en las restricciones léxicas circundantes.

La mayor desventaja es que el modelo representa el conocimiento lingüístico y por lo tanto debe especificarse totalmente para cada una de las palabras. Aunque fue muy reducida su aplicación, comparte conceptos con las gramáticas de dependencias, en ambas se considera que las palabras definen una información primordial que establece los enlaces con otras palabras. También comparte conceptos con las gramáticas de restricciones ya que cada experto de palabra especifica sus reglas de desambiguación.

GRAMÁTICAS DE CASO.

En las gramáticas de caso se considera que la estructura sintáctica es accesoria y sólo es importante si puede ayudar a la construcción de una representación en la cuál se especifiquen las diferencias de los roles semánticos. En estas gramáticas son primordiales los papeles o roles temáticos, por ejemplo, en las frases siguientes:

Francisco adora el triciclo.

Francisco comió un helado.

Francisco pescó un resfriado.

La estructura sintáctica de las tres frases es similar, se componen de un grupo nominal (*Francisco*), un verbo y un grupo nominal formado por un determinante y un sustantivo; sin embargo, en esta teoría, lo más importante es que hay un sentido muy diferente en las tres que no está relacionado con la estructura sintáctica. En la primera frase se expresa una edad, en la segunda una acción, y en la última frase se expresa un cambio.

Otra diferencia más puede notarse en las frases siguientes dónde se utiliza la preposición *con* de maneras muy diversas, para introducir una herramienta, un material, una manera en que se realiza la acción y el motivo.

Francisco construyó con madera de cedro.

Francisco lo construyó con un propósito específico.

Francisco lo construyó con precaución.

Francisco lo construyó con una herramienta de carpintería.

Entre las gramáticas de caso, Fillmore considera los diferentes roles semánticos [Fillmore, 68, 77], Grimes considera los roles específicos [Grimes, 75] y Schank estudia las dependencias conceptuales [Schank *et al*, 72].

[Fillmore, 68] sostiene que se puede identificar un conjunto de casos semánticos que permiten mostrar las relaciones de sentido que existen entre los sustantivos o grupos nominales y el verbo en frases simples. Su argumentación se funda en ejemplos como el siguiente:

Javier rompió la ventana con un martillo.

El martillo rompió la ventana.

La ventana se rompió.

En estas tres frases se describe la misma acción *romper*, y muestran que la misma función gramatical (sujeto) puede rellenarse con tres diferentes participantes. Así que tanto *Javier* como *el martillo* y *la ventana* son roles sintácticos iguales pero roles semánticos diferentes. Este fenómeno tiene una consecuencia importante, al establecer que las nociones de sujeto y objeto no son lo profundas que se requerirían para hacer esas diferencias. Por lo que la posición de la semántica se vuelve más importante que la sintáctica en esta teoría.

[Fillmore, 68] propuso el siguiente conjunto de roles semánticos: agente, instrumento, dativo, factitivo, lugar y objeto. [Fillmore, 71] modificó el conjunto a: agente, contra-agente, objeto, resultado, instrumento, fuente, meta y paciente. Aunque sus pretensiones fueron reducidas, su teoría tuvo una gran influencia en los formalismos gramaticales.

Otras teorías de casos buscaron un grado de abstracción mayor y trataron de identificar conceptos generales aclarando las relaciones que los diferentes casos tienen entre ellos. Por ejemplo, [Grimes, 72] tuvo la meta de encontrar un conjunto de casos más abstractos. Divide los casos en diferentes grupos: los roles de orientación (relativas al movimiento y a la posición), los roles de proceso (cambios de edad) y roles específicos. Una de las diferencias esenciales con la teoría de Fillmore es que un grupo nominal específico puede tener distintos roles semánticos.

[Schank *et al*, 72] intentó identificar nociones primitivas independientes del lenguaje, desarrolló un sistema de representación de sentidos denominadas

dependencias conceptuales, fundadas sobre las relaciones conceptuales entre los objetos y las acciones. Definió once primitivas en función de las cuales describió todas las acciones y precisó los roles conceptuales que pueden unir esas primitivas a los conceptos. Las diferencias fundamentales con las otras teorías son las siguientes:

- Un solo caso semántico puede unir entidades diferentes (por ejemplo, el caso *beneficiario* hace intervenir el donador y el receptor).
- Los casos ligados a una acción conceptual de base son todos obligatorios (si no son realizados en la frase, se pueden hacer inferencias para encontrarlos).
- Los casos semánticos unen las entidades conceptuales (y no los elementos sintácticos como GN, GP) a una de las once acciones conceptuales de base (y no a verbos auxiliares de superficie).

Las gramáticas de caso se han empleado con utilidad en muchas representaciones semánticas de lenguajes. Estas gramáticas tienen la ventaja única de permitir el análisis de frases no normalizadas o que no respetan la sintaxis correcta, sin embargo el problema principal es identificar un conjunto universal de casos. Aún con la representación más ambiciosa, de Schank, de identificar nociones primitivas independientes del lenguaje solamente se han empleado en dominios muy precisos [Schank et al, 72], [Schank, 80].

Esta noción de roles semánticos generalizados es compartida por los formalismos más recientes en el enfoque de constituyentes y con la misma dificultad de identificar un conjunto universal de roles.

Convergencia de los dos enfoques

Antes de presentar la convergencia de los dos enfoques presentados, exponemos una comparación de los formalismos presentados en cuanto a implementación y descripción de dependencias lejanas. Aunque aquí presentamos los formalismos más representativos en cada uno de ellos, existen otras variantes de los mismos por lo que generalizamos los nombres de los formalismos.

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 significante 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 (tema de la siguiente sección) 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.

En la siguiente tabla presentamos cómo se ha ido disminuyendo el número de reglas y transformaciones a expensas de la riqueza de información en el diccionario, y la aparición de restricciones e integración semántica. La marca X denota existencia, la marca — denota ausencia, y las otras marcas indican movimientos de incremento y reducción.

	Reglas CFG	Transf.	Diccio-nario	Restric-ciones.	Integra semántica	Estructura Múltiple	Estructura Comuni-cativa
GGT	X	X	—	—	—	—	—
ST	X	X	—	—	—	—	—
EST	X	X	X	—	—	—	—
GB	X	1	X	—	—	—	—
GPSG	X-	—	X	—	—	—	—
LFG	X--	—	X+	X	X	X	X
CG	X--	—	X++	X	X	—	—
HPSG	X---	—	X+++	X	X	X	—
DUG	—	—	X+++	X	X	—	—
MTT	—	—	X+++	X	X	X	X

-	inicio de reducción	+	concepción mejorada
--	reducción	++	importante
---	casi eliminación	+++	mayoría de la información

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 y 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 sicolingüística 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[16] 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 sicolingüista.

Esta observación, modifica el concepto de estructura de frase en la HPSG, donde la noción de estructura de frase se construye alrededor del concepto núcleo-*h* 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.), esta noción y su similitud con la MTT quedará de manifiesto en la siguiente sección dedicada a las valencias sintácticas.

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 aquéllas.

[16] Pivote con el sentido de álgebra de matrices.

1.2 VALENCIAS SINTÁCTICAS: ENFOQUES DIVERSOS

Las entradas léxicas en diccionarios manuales llevan una gran cantidad de información diferente acerca de los lexemas. Una pieza muy importante de información que algunos de los lexemas llevan es la información que algunos lingüistas llaman subcategorización. La información de subcategorización especifica la categoría del lexema, su número de argumentos, la categoría de cada argumento y usualmente la posición respecto al lexema, adicionalmente a veces se incluye también la información de las características como género, número, etc.

El ejemplo más simple de subcategorización es la diferencia entre un verbo transitivo y uno intransitivo; un verbo transitivo debe tener un objeto a fin de ser gramatical, por ejemplo:

María ablanda la carne.

**María ablanda.*

Y un verbo intransitivo no puede tener un objeto, por ejemplo:

María cojea.

**María cojea una pierna.*

En el ejemplo previo, *ablandar* es un verbo y debe aparecer inmediatamente precediendo un grupo nominal GN (*la carne*). Se dice que ese verbo *subcategoriza* un GN. A partir de esta clasificación simple, transitivos e intransitivos, se amplía la información para considerar todos los casos posibles, por ejemplo la doble transitividad [Cano, 87] considera que el verbo subcategoriza dos complementos.

En el procesamiento lingüístico de textos por computadora, básicamente la subcategorización se refiere al número de argumentos y la categoría de cada argumento pero la forma de definir cuáles son y cómo se representan los argumentos subcategorizados por un lexema dado ha diferido en los diversos formalismos en los dos enfoques considerados en el análisis sintáctico. En el enfoque de dependencias,

donde se emplean muchos de los términos de la gramática tradicional, para nombrar esta información se emplea el término valencia sintáctica que nosotros seguimos en el título y en algunos subtítulos de esta sección.

En el enfoque de constituyentes, la subcategorización se representa en términos sintácticos, es decir, por su estructura y parte del habla. Los verbos pueden subcategorizar diferentes tipos, no solamente los GN, por ejemplo, el verbo *dar* subcategoriza un grupo nominal (GN) y un grupo preposicional (GP), en ese orden: *Juan da un libro a María.*

Aunque, desde el punto de vista de este enfoque, la subcategorización se describe de una manera más fija, contrasta con las colocaciones. Las colocaciones describen los contextos locales, que son importantes ya de una manera preferencial o estadística, en la frase. Por ejemplo, en el proyecto DECIDE para construcción de recursos: diccionarios y corpus principalmente, [DECIDE, 96], se considera la información de subcategorización (*subcat*) como una lista con frecuencias de aparición de diferentes palabras unidas a la palabra seleccionada, en un corpus. En este diccionario, incluso aparecen las combinaciones con una sola ocurrencia, que solamente tiene un significado estadístico y que no representan la realización de un complemento.

En el enfoque de constituyentes o gramáticas de frase, la selección semántica no es una condición ni suficiente ni necesaria para la subcategorización. Así que la mayoría de estas teorías lingüísticas incluyen en el marco de subcategorización predicados[17] o frases cuya ocurrencia es obligatoria en el contexto local de la frase del predicado aunque no sean seleccionados semánticamente por él.

Dentro del enfoque de constituyentes presentamos, en esta sección, la descripción de las valencias sintácticas para los formalismos GB, GPSG, LFG, CG y HPSG.

Las teorías lingüísticas basadas en dependencias incluyen, en la información de las valencias sintácticas, las frases cuya ocurrencia es obligatoria en el contexto semántico del verbo. Adicionalmente, algunos formalismos, consideran los complementos circunstanciales, con una clara distinción entre ellos y los especificados semánticamente. Este razonamiento se basa en separar las alternaciones de valencias, específicas de cada lexema, y los complementos circunstanciales, comunes a distintos lexemas.

Tanto en la WG como en la MTT las valencias sintácticas describen únicamente las frases cuya ocurrencia es obligatoria en el contexto semántico del verbo. En cambio, la DUG y la Gramática Funcional de Dependencias (FDG, *Functional Dependency Grammar*, en inglés) [Tapanainen *et al*, 97] adicionalmente

[17] Los predicados manifiestan lo que se dice del sujeto en la oración, por lo que la mayoría de los formalismos del enfoque de constituyentes no consideran el sujeto dentro de la valencia sintáctica.

describen los predicados circunstanciales. Dentro del enfoque de dependencias, presentamos la descripción de las valencias sintácticas para los formalismos DUG y MTT.

Así que, en general, la valencia sintáctica o subcategorización concierne con la especificación de frases que son preponderantes al contexto del verbo porque son seleccionadas por el lexema, sintácticamente o semánticamente o ambas. Aunque todas las teorías lingüísticas tienen medios para expresar los aspectos sintácticos, y morfosintácticos, de subcategorización, la referencia directa a la selección semántica puede expresarse únicamente en aquellos formalismos que incluyen un nivel de representación semántica.

Desde el punto de vista del procesamiento lingüístico de textos, la especificación de la estructura de las valencias sintácticas es necesaria para codificar la información concerniente al contexto y al orden de palabras a fin de limitar el análisis y la generación del lenguaje natural, este argumento se explicará más adelante. La complejidad resulta por el aspecto multidimensional de la estructura de las valencias sintácticas, porque la subcategorización involucra referencia a diversos niveles de descripción gramatical, aspectos morfológicos, sintácticos y semánticos de la especificación de las palabras, y también por la interfase entre estos niveles de descripción gramatical.

Se ha puesto una gran atención a esta información en los diccionarios computacionales como COMLEX [Grishman *et al*, 94] no solamente para verbos sino para adjetivos y sustantivos que llevan complementos. En el procesamiento lingüístico de textos, esta información ayuda a establecer las combinaciones posibles de los complementos en la oración. Pero también tienen importancia relevante para la traducción automática, por ejemplo [Fabre, 96] estudió las relaciones predicativas de sustantivos para interpretar compuestos nominales en francés e inglés.

Las teorías lingüísticas difieren en la cantidad de información que proveen en la valencia sintáctica de un verbo. Esto se debe, en su mayoría, a las diferentes tendencias al usar principios y reglas sintácticas para expresar generalizaciones lingüísticas, con el consecuente cambio de énfasis más lejano o más próximo a la especificación léxica. En esta sección presentamos una revisión de diversos enfoques adoptados en las teorías lingüísticas y a continuación un análisis de ellos.

Subcategorización en GB

En el desarrollo de la GB se percataron de la gran redundancia de información en las reglas de estructura de frase y en los marcos de subcategorización. Por ejemplo, la información de que un verbo transitivo va seguido de un objeto tipo GN estaba codificada tanto en la regla que expande el GV como en el marco de subcategorización del verbo. La GB movió esta información a los marcos de subcategorización de los núcleos-*h*. La razón para hacer esto es que cada verbo

selecciona-*c* *(c* por categoría) un cierto subconjunto del rango de proyecciones máximas.

La teoría de la X-barra presenta la idea de que se encuentran patrones similares dentro de cada una de las estructuras internas de diferentes frases en un lenguaje. Por ejemplo, tanto el verbo como las preposiciones preceden a su objeto. El núcleo-*h* de una unidad lingüística es esa parte de la unidad que da su carácter esencial. Así, el núcleo-*h* de un GN es el sustantivo, similarmente, un verbo es el núcleo-*h* de un GV, y así sucesivamente.

En este formalismo, la frase es una proyección del núcleo. Se consideran dos niveles de proyección. Por ejemplo, en el nivel más bajo el núcleo léxico y los argumentos (constituyentes a los cuales subcategoriza el núcleo) denotados con una barra o un apóstrofo ($\overline{\mathbf{N}}$, N'), y en el siguiente nivel esa misma estructura con modificadores y especificadores, denotados con dos barras o dos apóstrofos ($\overline{\overline{\mathbf{N}}}$, N''). Esta última es la máxima proyección, donde N'' es igual que GN, V'' igual a GV, etc.

Un ejemplo de modificadores y especificadores son los adjetivos y artículos para N'. No hay duda de que cualquier proyección máxima (es decir, GA, GN, GP, O', o GV) puede ser el argumento de un núcleo-*h*, en principio, aunque típicamente, núcleos-*h* diferentes seleccionan elementos diferentes del conjunto de proyecciones máximas como sus argumentos. El verbo *ablandar* selecciona GN, *decir* selecciona O' (como en *dijo que la carne estaba lista*), etc.

De estas nociones se ve como la información de subcategorización limita el análisis y la generación de lenguaje natural. La subcategorización se usa como un filtro en el análisis y en la generación de estructuras de frase, en el sentido siguiente: si tratamos, por ejemplo, de hacer la inserción léxica de *ablandar* en una estructura donde es hermana izquierda de una O', esa estructura con ese núcleo-*h* se descartará, porque su subcategorización requiere un GN.

En la GB la relación indirecta entre el verbo y su sujeto es un aspecto crucial de la teoría total y está presente en todos los análisis. El sujeto, en inglés, no aparece como hermano del núcleo-*h* del GV y por lo tanto no puede ser subcategorizado por ese núcleo-*h*. El dominio de subcategorización está limitado al dominio de la proyección máxima que contiene el núcleo-*h*, y es realmente esta noción de dominio dentro de la proyección máxima, en lugar de la noción de ser hermana, la que es importante en esta teoría. El sujeto no está dentro del dominio del verbo ya que la proyección máxima del verbo es GV. Esto resulta en las diferencias tanto del comportamiento sintáctico del sujeto y de los complementos (que no son sujetos) como en el hecho de que el sujeto es externo al GV (ver Figura 10). Así, los complementos que no son sujetos son los únicos que pueden subcategorizarse en este formalismo.

El sujeto es el GN inmediatamente dominado por O, y el objeto es el GN inmediatamente dominado por el GV. En la GB, esto se representa comúnmente por

las notaciones [GN, S] y [GN, GV] respectivamente. El uso de los términos sujeto y objeto en este formalismo son las abreviaturas de esas definiciones estructurales. Desde este punto de vista, el objeto de la estructura-*d* puede volverse en el sujeto de la estructura-*s* en la construcción pasiva.

La subcategorización en la GB se describe en un nivel de descripción sintáctica donde los argumentos de un predicado se juntan en un conjunto donde cada elemento corresponde a un papel temático indexado [Williams, 80]. Dentro de la estructura de argumentos de un predicado puede haber una posición distinguida que funciona como el *papel temático del núcleo-h* de la estructura de argumentos como una totalidad. Este papel temático se denota como el *argumento externo* ya que puede ser asignado solamente afuera de la proyección máxima de su predicado.

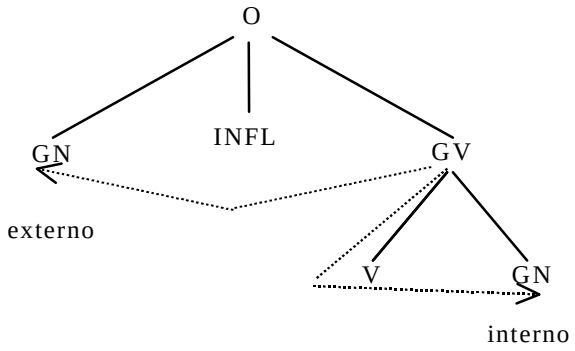


Figura 10. Relación indirecta entre sujeto y objeto

En versiones posteriores de la GB [Chomsky, 86], a diferencia de la mayoría de las otras teorías gramaticales, las frases se asumen como las proyecciones máximas de la frase con inflexión, la que introduce la morfología verbal (por ejemplo, tiempo y aspecto). En la Figura 10, INFL es la inflexión.

La descripción en la Figura 11, corresponde a [Sells, 85], la subcategorización (selección categorial) en paréntesis angulares y la estructura de argumentos (selección semántica) en paréntesis, donde el argumento externo está subrayado siguiendo la notación de [Williams, 81]. La información de los papeles temáticos restantes, es decir, de los *argumentos internos,* está disponible únicamente dentro de la primera proyección del predicado.

La realización sintáctica de los papeles temáticos en la estructura del argumento se limita y asegura por el Principio de Proyección y por el Criterio-Theta, que a continuación se presentan.

- *Principio de Proyección.* Las representaciones en cada nivel sintáctico (es decir la forma lógica y las estructuras -*d* y -*s*) se proyectan desde el

diccionario, siguiendo las propiedades de subcategorización de los elementos léxicos.

- *Criterio*-θ. Cada argumento sostiene uno y sólo un papel-θ, y cada papel-θ está asignado a uno y solamente un argumento.

El criterio-θ dice en forma simple que el significado de un predicado determina qué argumentos gramaticales tendrá. El principio de proyección garantiza que la estructura determinada por el significado léxico del núcleo-*h* no sea alterada en forma esencial.

También hay un principio que relaciona la subcategorización y la asignación de papeles-θ o papeles temáticos (comúnmente llamado *marcado*-***T***). La subcategorización se relaciona a posiciones en un arreglo y el *marcado*-***T*** al contenido léxico dominado por esa posición. Si α subcategoriza la posición ocupada por β, entonces α marca-θ a β.

ablandar, V, <GN> (Agente, Tema)

dar, V, <GN, GP> (Agente, Tema, Meta)

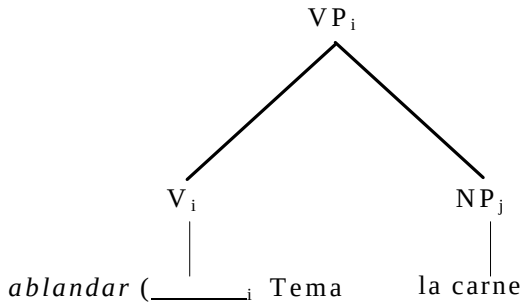


Figura 11. Subcategorización y papeles temáticos

Como la subcategorización está relacionada a posiciones, debe codificarse algún tipo de posición de argumento temático para el sujeto, en la entrada léxica del verbo. En [Chomsky, 86] se asume que la selección categorial (selección-*c*) puede derivarse como la Realización Estructural Canónica (CSR) de su categoría semántica. Por ejemplo, la CSR (rol paciente) es un grupo nominal. Consecuentemente, solamente la selección semántica (selección-*s*) necesita expresarse en el diccionario.

En el enfoque de constituyentes, la GB dentro de ella, también se consideran los predicados no seleccionados semánticamente, como los casos de complementos de verbos cuyo sujeto es pleonástico (*extraposition*, en inglés), verbos que se denominan

"de ascensión"[18] (*raising verbs* en inglés), por ejemplo *seem*, y verbos que contrastan con estos últimos, los denominados verbos de control (*control verbs*[19] o *equi*[20] *verbs*, en inglés). Por ejemplo:

- Sujeto pleonástico: *It annoys people that dogs bark*. (Molesta a la gente que los perros ladren). El pronombre neutro *it* representa *dogs bark* (los perros ladran), el sujeto del verbo *annoy* (molestar). Sintácticamente existen dos argumentos correspondientes al mismo argumento semántico. El nombre "extraposición" viene del análisis transformacional, teniendo la frase *that dogs bark annoys people* se proponía cambiar de posición la cláusula *that dogs bark* al final de la frase e insertando el pronombre vacío *it*. En español no se requiere esa inserción, por ejemplo: Q*ue no se le atienda a tiempo molesta a la gente* y *Molesta a la gente que no se le atienda a tiempo*.

- Verbos de ascensión: *Mary seems to be happy*. (María parece ser feliz.) y *I expected Mary to be happy* (Yo esperaba que María fuera feliz.). Se considera que cada verbo tiene un sujeto, incluso el infinitivo. En la primera frase, el sujeto del primer verbo (sujeto de ascensión, *subject raising*, en inglés) es transparente en cuanto a que también es sujeto del segundo verbo (María parece, María es feliz). En la segunda frase el objeto del primer verbo (objeto de ascensión, *object raising*, en inglés) es el sujeto del segundo verbo (esperaba que María, María fuera). En español existen muchos verbos que introducen otros verbos, ya sea directamente como *querer*, *poder*, o mediante una preposición como *ponerse a bailar*, *deben de cantar*, etc. Un estudio de verbos españoles, con este punto de vista, se presenta en [Lamiroy, 94].

La teoría de control en la GB maneja sintácticamente los verbos *equi*. En estos verbos, el sujeto de verbos no finitos, es decir, de grupos verbales en infinitivo, se representa estructuralmente como la categoría vacía PRO cuya relación a su controlador está regulada por la Teoría del Ligamento en términos del comando-*c*, que expresa algo así como la noción de esa subparte de un árbol para la cual una categoría determinada α no es inferior jerárquicamente.

María $_i$ *intenta [PRO* $_i$ *dormir]*

Esto implica que la subcategorización verbal, de cláusula, se expresa siempre en términos de oraciones en lugar de hacerlo en términos de grupos verbales.

---

[18] Donde solamente la posición controlada es temática

[19] El controlador y el controlado son ambos temáticos, la predisposición de control se especifica léxicamente.

[20] Verbos de control son lo mismo que *equi-NP deletion*, que se abrevia *equi*.

Las dependencias verbales que emergen en las construcciones expletivas[21] y de sujeto de ascensión se manejan también sintácticamente. Por ejemplo, un verbo de ascensión como *seem (parecer)* subcategoriza una frase pero no tiene argumento externo. Existen dos casos cuando se subcategoriza una cláusula:

- Si la cláusula subcategorizada no es finita, el sujeto se mueve a una posición de sujeto en el arreglo para satisfacer el Filtro de Caso[22] puesto que solamente un GV con marca de tiempo puede asignar caso nominativo a su sujeto. Por ejemplo: $Juan_i$ *parece* $[t_i$ *dormir]* donde t_i es la huella del sujeto i.
- Si la cláusula subcategorizada es finita, por ejemplo en *It seems that John sleeps* (parece que Juan duerme), el elemento pleonástico *it* se inserta en la posición sujeto del arreglo para satisfacer el Principio de Proyección Extendida que además del Principio de Proyección anterior requiere que todas las cláusulas tengan sujeto.

Por último, las construcciones con objeto de ascensión también se consideran como si involucraran subcategorización de oraciones. Un verbo como *believe* subcategoriza una frase de infinitivo a cuyo sujeto se le asigna caso por el verbo en el arreglo, a través de límites de oraciones, como en *Mary believes* $[_S$ *John to be intelligent]* que es una ocurrencia descrita como *marcado de caso excepcional*, en [Chomsky, 86].

Subcategorización en GPSG

La GPSG hace uso de características sintácticas, de entre ellas, dos ejemplos son las siguientes: una para mostrar el POS y otra para mostrar el nivel (palabras, grupo de palabras, frase). Además desarrolla una teoría apropiada de características, expresándolas mediante pares de atributos y valores. No solamente se consideran como atributos las categorías como número, caso y persona, sino también el nivel, esto es influencia de la teoría X-barra, y también con la misma interpretación.

En la GPSG se emplea un atributo para la subcategorización, llamado SUBCAT, y se asigna un valor único a cada posible marco en el cual pueda ocurrir una categoría de nivel cero. SUBCAT es una característica del núcleo-*h*, es decir, de HEAD. Por ejemplo, si la entrada léxica *comer* sólo dice que es un verbo transitivo, es decir, [SUBCAT TRANS], entonces el hecho de que los verbos transitivos, y sólo ellos, ocurran con un nodo hermano GN puede establecerse mediante una regla ID como:

---

[21] En las construcciones expletivas, el argumento pleonástico se encuentra en la posición ya sea de sujeto u objeto.

[22] El *Filtro de Caso* especifica que a cada GN léxico debe asignársele caso.

V1 ⟶ V0 [SUBCAT TRANS], NP

donde V0 es el verbo, V1 es el grupo verbal y V2 es la máxima proyección. La GPSG comparte, con la GB, el análisis de que la máxima proyección del verbo es la oración. Una categoría puede dominar un elemento léxico si y sólo si la categoría es consistente con la entrada léxica de ese elemento. Así que sólo un verbo que sea TRANS, como *comer*, puede ocurrir bajo V0 [TRANS] y uno intransitivo como *cojear* no podrá.

Realmente los verbos no tienen un marco de subcategorización, sino que tienen una indicación que apunta al tipo de estructura en la que aparece. Para considerar todos los posibles tipos, GPSG utiliza números enteros como valores de SUBCAT, y los incluye en las entradas léxicas y en las reglas ID, correspondiendo a las estructuras posibles. A continuación se presentan unos ejemplos:

V1 ⟶ V0[1]

V1 ⟶ V0[6], NP, PP

cojear: V0[1]

dar: V0[6]

La GPSG considera posible que un verbo tenga múltiples subcategorizaciones. Cada estructura de subcategorización corresponderá con una entrada léxica separada pero relacionada al lexema. En la GPSG existen postulados de sentido que imponen relaciones sistemáticas entre los sentidos de verbos homónimos. Estos postulados de sentido son precisamente postulados semánticos, y es en términos semánticos que la GPSG captura el hecho de múltiples subcategorizaciones.

Un problema evidente de esta teoría es que implica un gran número de reglas ID. Algo de la redundancia en ellas se elimina mediante el uso de postulados LP separados (por ejemplo, para dictar el orden de los nodos hermanos en un subárbol), y otra parte se elimina por los principios de características. Pero la esencia de la objeción permanece.

Los objetos sintácticos como sujeto y objeto no se consideran nociones primitivas en la GPSG, sino que se definen en términos de otras primitivas de la teoría. En la GPSG, siguiendo a [Dowty, 82] esas relaciones se definen en términos de la estructura semántica, es decir, en la estructura función-argumento de la semántica. Por ejemplo, un verbo transitivo como *buscar* requiere dos argumentos. El sujeto se define, sólo semánticamente, como el último argumento, el objeto es el siguiente del último, etc.

La diferencia entre verbos de ascensión y *equi* se define en la subcategorización de los verbos, es decir, en las reglas-ID que producen los nodos que los dominan inmediatamente en las estructuras sintácticas. Por ejemplo:

VP ⟶ H[15], VP[INF, +NORM] *try*

VP ⟶ H[16], VP[INF] *seem*

Donde +NORM es la abreviatura de AGR NP[NFORM NORM], que establece la concordancia del grupo nominal. Mientras para el verbo *seem* se permite cualquier sujeto, para el verbo *try* es necesario que el sujeto mediante concordancia (NORM) no pueda ser ni *it* ni *there*. Una complejidad se presenta al establecer los valores de omisión para *seem.* VFORM es una característica de HEAD que distingue partes del paradigma verbal: FIN (finito), INF (infinitivo), BSE (forma base), PAS (pasiva), etc.

En la GPSG, el núcleo-*h* sólo puede subcategorizar sus hermanas, por lo que los sujetos no se subcategorizan. Realmente no hay subcategorización para el sujeto, aunque este hecho a veces es dudoso porque la existencia de la característica AGR para manejar la concordancia entre sujeto y verbo, tiene el efecto como de permitir la subcategorización para los sujetos.

Subcategorización en LFG

La subcategorización en la LFG, como en otras gramáticas de constituyentes, se basa en una representación sintáctica de la estructura de los argumentos del predicado. Pero en la LFG, la noción de función gramatical ocupa un papel central para determinar cuáles argumentos, seleccionados semánticamente por un predicado, están realizados semánticamente y cómo.

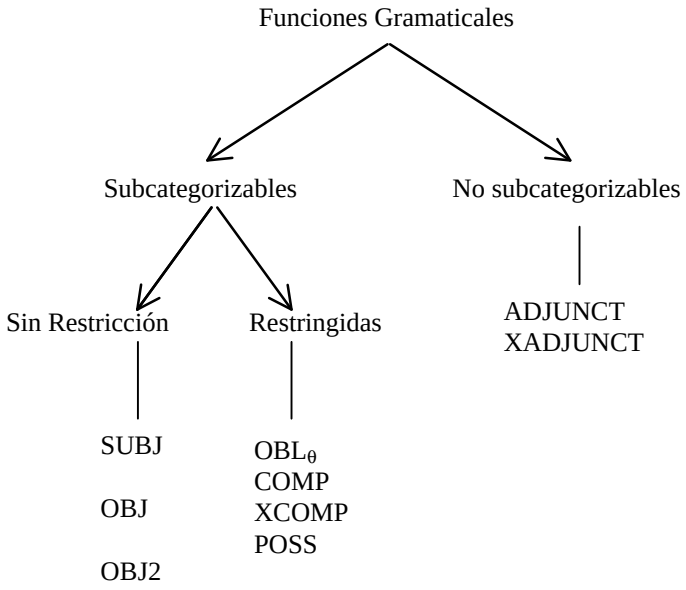
En [Bresnan, 82], las funciones gramaticales se definen como primitivas sintácticas universales de la gramática y se clasifican de acuerdo a dos parámetros principales: la habilidad de subcategorizar y la restricción semántica. Las funciones subcategorizables que pueden asignarse a los argumentos de los lexemas son los sujetos, los objetos y los complementos de los grupos verbales de la oración. Las funciones que no son subcategorizables corresponden a frases adjuntas que no pueden asociarse con los argumentos de los lexemas.

Existen otras funciones como Tópico y Foco que se asignan a las frases desplazadas, como en la topicalización, las preguntas y las cláusulas relativas. Se considera que la habilidad de subcategorizar de estas dos funciones está sujeta a variación lingüística, ya que es posible que exista en algunos lenguajes y en otros no.

En la LFG las funciones que se pueden subcategorizar difieren con respecto al rango de tipos de argumentos con los cuales pueden asociarse [Kaplan, 94], y se dividen en restringidas y sin restricción:

- Las funciones gramaticales *semánticamente no restringidas* no están ligadas de una manera inherente a las restricciones específicas de selección. Por ejemplo, la función sujeto, que puede realizar argumentos no temáticos como el sujeto *it* de *seem*; o aunque los sujetos son a menudo agentes, también pueden ser tema, como en la pasiva.
- Las funciones gramaticales *restringidas semánticamente* son las que están más íntimamente ligadas a la semántica, es decir, solamente pueden ponerse por pares con argumentos de tipos semánticos específicos. Por ejemplo, las funciones oblicuas (objeto directo, objeto indirecto) [Rappaport, 83], que siempre son temáticas, es decir, que nunca se asocian con elementos pleonásticos. En español, sí se presenta la duplicación de objetos como se verá en la sección 2.6.

En la siguiente figura se presenta la clasificación general de las funciones gramaticales y más adelante se describen individualmente.



OBL_θ significa oblicuo; POSS es el genitivo prenominal, como el caso en inglés de *professor's knowledge* (conocimiento del profesor).

También los complementos y los adjuntos se clasifican, en funciones cerradas o abiertas. Cerradas significa que están completas, tienen sus propios controladores, y abiertos lo opuesto, requieren antecedentes. En los ejemplos de complemento cerrado

y de función adjunta cerrada (COMP, ADJ) los GN subrayados son los controladores.

- Complementos. Los complementos cerrados son los COMP y los abiertos XCOMP.

  Beto cree [que María es honesta]$_{COMP}$

  Beto intenta [ser un buen médico]$_{XCOMP}$

- Adjuntos. Los adjuntos cerrados son los ADJUNCT y los abiertos XADJUNCT.

  [Beto empezaba a alegar]$_{ADJ}$, María salió despavorida.

  [Aún estando enojado]$_{XADJ}$ Beto comió tranquilamente.

Los objetos sintácticos son asociaciones de funciones gramaticales con papeles temáticos o con valores que no son temáticos. Estas asociaciones se codifican en el diccionario, donde cada verbo está representado como un lexema que consiste de una estructura de argumentos del predicado y una asignación de función gramatical. Por ejemplo:

Estructura de argumento de predicado	romper	<*agente*, *tema*>
Asignación de función gramatical		((SUJ), (OBJ))

Donde la estructura de argumentos del predicado de un lexema es una lista de los argumentos para los cuales existen restricciones de selección. La asignación de función gramatical de una forma léxica es una lista de sus funciones subcategorizadas sintácticamente.

La asignación de funciones gramaticales se sujeta a un número de condiciones universales. Por ejemplo, todos los predicados univalentes se asignan a SUJ, y todos los predicados bivalentes se asignan a un SUJ y a un OBJ. Una condición muy importante sobre la asignación de función gramatical es la Biunicidad de las Asignaciones Función-Argumento [Bresnan, 82] que establece una relación uno a uno entre argumentos y funciones gramaticales dentro de la estructura predicado-argumento de un lexema.

Esas listas de asignación de función gramatical sirven como marcos de subcategorización. La subcategorización se revisa en la estructura funcional mediante dos condiciones: *Completeness* y *Coherence* [Kaplan & Bresnan, 82]:

- La completitud asegura que todos los argumentos subcategorizados estén presentes en la estructura funcional, es decir, que no haya menos argumentos. Por ejemplo, descarta frases como **Juan compra*, **seems.*

- La coherencia restringe la ocurrencia de funciones gramaticales subcategorizables a las listadas en la forma léxica del verbo, es decir, que no haya argumentos de más. Por ejemplo, descarta frases como **Juan cojea Memo*.

Finalmente, el *control funcional* maneja léxicamente los verbos de control y de ascensión con referencia a funciones gramaticales. Por ejemplo, el control del sujeto con ambos tipos, de ascensión y de control, se establece en el diccionario en las partes relevantes de las entradas léxicas como en los siguientes.

V (↑PRED) = '*seem* < (↑XCOMP) > (↑SUJ)'

(↑XCOMP SUJ) = (↑SUJ)

V (↑PRED) = '*try* < (↑XCOMP) (↑SUJ)>'

(↑XCOMP SUJ) = (↑SUJ)

En la descripción del verbo *try* se especifica que el sujeto es temático. Ya que el control se trata léxicamente y las categorías no vacías se usan para unir el sujeto complemento, se obtiene que ambos verbos (de ascensión y control) subcategorizan grupos verbales en lugar de oraciones, como se considera en la GB.

En los trabajos de [Bresnan & Kanerva, 88] y [Bresnan & Moshi, 89], entre otros, se revisó la teoría de los objetos sintácticos. Los objetos sintácticos como SUJ, OBL, etc., pasaron de especificaciones atómicas a definiciones en términos de características funcionales más primitivas. La teoría resultante, la Teoría Léxica de Mapeo, consiste de cuatro componentes básicos

- Jerarquía de papeles léxicos. La jerarquía incluye los siguientes papeles en orden descendente: agente, beneficiario y maleficiario, receptor y experimentador, instrumental, paciente y tema, locativo, motivo; se crea una jerarquía temática universal en base a ellos.

- Funciones sintácticas no compuestas. Las funciones sintácticas se descomponen de acuerdo a las características [± r], temáticamente restringidos o sin restricción, y [± o], objetivo o no, por ejemplo:

$$\begin{bmatrix} -r \\ -o \end{bmatrix} \text{ SUJ} \qquad \begin{bmatrix} -r \\ +o \end{bmatrix} \text{ OBJ}$$

  Individualmente, cada valor de las dos características define una función gramatical parcialmente especificada, por ejemplo:

$$[-r] \text{ SUJ/OBJ} \qquad [+o] \text{ OBJ/OBJ}_q$$

- Principios de mapeo léxico. Los papeles semánticos se asocian con funciones gramaticales especificadas parcialmente de acuerdo a los Principios de Mapeo Léxico: clasificaciones de roles intrínsecos, clasificaciones de roles morfoléxicos y clasificaciones de roles por omisión.

- Condiciones de buena formación. Después de que los principios de mapeo se han aplicado, cualquier función gramatical restante no bien especificada está totalmente instanciada. Esta instanciación es libre tanto como se

observen los principios de Biunicidad y de Condición de sujeto. El primero establece que dentro de la estructura de un predicado-argumento de una forma léxica hay una relación de uno a uno entre funciones gramaticales y argumentos. La condición sujeto establece que cada forma léxica debe tener un sujeto.

Como ejemplo de la aplicación de esta Teoría léxica de mapeo se presenta el tratamiento de la forma pasiva, de [Bresnan & Kanerva, 88]. Para el verbo *buscar*, antes de la conversión a pasiva, los papeles de agente y tema del verbo están intrínsecamente asociados con funciones gramaticales parcialmente especificadas, como se muestra a continuación:

$$\begin{array}{lcccc} buscar & \langle & agente & tema & \rangle \\ & & | & | & \\ & & [-o] & [-r] & \end{array}$$

La regla pasiva introduce la especificación funcional [+r], es decir, restringida temáticamente, para el papel superior de una forma léxica. Cuando la pasiva se aplica a la estructura de argumentos de predicado para el verbo *buscar*, el papel del agente adquiere la especificación [+r] que en conjunto con [-o] define una función oblicua. El argumento agente de un verbo pasivo se realiza como un complemento oblicuo, mientras el tema puede ser sujeto u objeto. Las restricciones de buena formación inducidas por la condición de sujeto requieren que se elija la opción sujeto en este caso. A continuación el ejemplo del proceso descrito, con una representación esquemática:

$$\begin{array}{lcccc} & buscar & \langle & agente & tema & \rangle \\ & & & | & | & \\ intrínseco: & & & [-o] & [-r] & \\ pasiva: & buscado & & [+r] & & \end{array}$$

$$\begin{array}{lcc} & \text{OBL}_{\hat{q}} & \text{OBJ/SUJ} \\ condición\ buena\ formación: & \text{OBL}_{\hat{q}} & \text{SUJ} \end{array}$$

Subcategorización en CG

En la aplicación de la Gramática Categorial al estudio de lenguajes naturales se ha supuesto una colección universal de esquemas de estructura de frase, también se ha supuesto que la estructura sintáctica determina la semántica funcional, de tipo composicional. De lo anterior deriva que todas las generalizaciones de lenguaje específico deben determinarse léxicamente, por lo que una vez establecido el diccionario para el lenguaje pueden aplicarse las reglas universales de combinación sintáctica y semántica.

En el proyecto ACQUILEX [Sanfilippo, 93] se aplicó la Gramática Categorial

de Unificación y en base a la descripción del marco ahí empleado se presenta a continuación la subcategorización. Una descripción más amplia de las estructuras de grupos verbales para el inglés se encuentra en [Carpenter, 95].

La información de subcategorización en esta aproximación se encuentra dentro de la estructura de signo. Los signos están formados por una conjunción de pares atributo–valor de información ortográfica (ORTH), sintáctica (CAT) y semántica (SEM). Las palabras y las frases se representan como estructuras de características, con tipos, mediante signos.

[ORTH: orth

CAT: cat

SEM: sem]

El atributo *categoría* de un signo puede ser básico o complejo:

- Las categorías básicas son las estructuras binarias de características que consisten de un tipo categoría, y una serie de pares atributo valor que codifican información morfosintáctica (cuando es necesaria). Los tipos *cat* básicos que se emplean son: sustantivo (n), grupo nominal (np) y oración (sent).

  [CAT–TYPE: cat–type

  M–FEATS: m–feats]

  Por simplicidad, se abrevian como: cat–type [m–feats]

- Las categorías complejas se definen recursivamente, dejando que el tipo *cat* instancie una estructura de características con los siguientes atributos: resultado (RES) que puede tomar como valor una categoría básica o una compleja, activo (ACT) que es de tipo signo, y direccción (DIR) que codifica el orden de combinación, relativo a la parte activa del signo (por ejemplo: hacia adelante o hacia atrás).

  [RES: cat

  DIR: dir

  ACT: sign]

En los verbos, la parte activa de la estructura de categorías codifica las propiedades de subcategorización. Por ejemplo, sujeto (nom) y objeto (acc) en verbos transitivos:

[ORTH: < ablandar>

CAT: [RES: [RES: sent

ACT: [np–signo

CAT: nom]]

ACT: [np–sign

CAT: np [acc]]]]

La información semántica de un signo es una fórmula. Esta fórmula consiste de:

- Un índice (IND) que es una entidad que provee información referida a un tipo ontológico. El índice "e" indica eventualidades, "o, x, y, z" objetos individuales
- Un predicado (PRED), el argumento de un predicado puede ser una entidad o una fórmula.
- Al menos un argumento (ARG1) que puede ser a su vez una entidad o una fórmula, subsumidas por *sem*.

[IND: entidad

PRED: pred

ARG1: sem]

Por ejemplo, la estructura de características:

[IND: [1] x

PRED: carne

ARG1: [1]]

donde [1] indica valores reentrantes. Por simplicidad las fórmulas se presentan en forma lineal, pueden abreviarse como *<x1> carne (x1)* donde x1 es una variable con nombre.

La clasificación de tipos de subcategorización involucra la definición de las *estructuras semánticas predicado-argumento*, de las *estructuras de categorías*, y de los *signos de los verbos*. Así que primero presentamos las descripciones de estos tres tipos de estructuras con los ejemplos únicos necesarios para mostrar, al final, la subcategorización completa de verbos de dos y tres argumentos.

Para describir las *estructuras semánticas predicado-argumento*, siguieron la clasificación de [Dowty, 89]. Así que el contenido semántico de las relaciones temáticas se expresa en términos de conceptos de grupos prototípicos: los roles proto-agente (*p-agt*) y los roles proto-paciente (*p-pat*), determinados para cada elección de predicado. [Sanfilippo & Poznanski, 92] además de formalizar los proto-roles como superconjuntos de grupos específicos de componentes significantes que son instrumentos en la identificación de clases semánticas de verbos, introdujeron adicionalmente dos conceptos:

- Un tercer proto-rol, *prep*, para argumentos preposicionales. Estos *prep* se consideran semánticamente restringidos, empleando los términos de la LFG.
- Los predicados sin contenido (no-θ) para caracterizar la relación entre un GN pleonástico y su verbo rector.

Los verbos se caracterizan como propiedades de eventualidades, y los roles temáticos son relaciones entre eventualidades e individuos, por ejemplo, p-agt(e1, x). Una clasificación semántica primaria de los tipos de verbos se obtiene en términos de la aridad del argumento, es decir, del número de argumentos. Las diferencias adicionales se hacen según qué tipo de argumentos verbales se codifican, por ejemplo: proto-agente, proto-paciente, preposicional oblicuo/indirecto, preposicional de objeto, no - temático, pleonástico, predicativo (como *xcomp*), oracional (como *comp*).

A continuación, las principales estructuras semánticas de verbos, con ejemplos:

STRICT–INTRANS–SEM Intransitivos estrictos. *Juan* (proto-agente) *cojea*

<e1> and (<e1> pred (e1), <e1> p–agt (e1, x))

STRICT–TRANS–SEM Transitivos estrictos. *Juan* (p-ag) *bebe una cerveza* (p-pat)

<e1> and (<e1> pred (e1), <e1> and (<e1>p–agt(e1,x), <e1>p–pat(e1,y)))

OBL–TRANS/DITRANS–SEM Ditransitivos: *dar*

Transitivos con complemento oblicuo. *Juan da un libro a María.*

<e1> and (<e1> pred (e1), <e1> and (<e1>p–agt(e1,x),

<e1> and (<e1>p–pat(e1,y), <e1> prep (e1,y))))

P–AGT–SUJ–INTRANS–XCOMP/COMP–SEM Intransitivos con sujeto temático y complemento tipo cláusula (representada por *verb-sem*). *Juan intentó venir* y *Juan pensó que María vendría.*

<e1> and (<e1> pred(e1), <e1> and (<e1>p–agt(e1,x), verb– sem))

Las *estructuras de categoría* se distinguen de acuerdo a los valores de las características RES y CAT. Por ejemplo, el CAT de intransitivos estrictos establece que el resultado es una categoría básica de tipo *sent* y la parte activa es un grupo nominal, es decir, solamente hay selección de sujeto. A partir de tipos básicos se van construyendo tipos más complejos de categoría. Los transitivos estrictos emplean la categoría de intransitivo estricto, dando adicionalmente la categoría acusativo al objeto.

STRICT–INTRANS–CAT STRICT–TRANS–CAT

[RES: sent [RES: strict–intrans–cat

ACT: np–sign] ACT: [np–sign

CAT: np[acc]]]

Las restricciones morfosintácticas se codifican en signos seleccionados (activos). Por ejemplo, en la definición de la categoría ditransitiva el argumento extremo tiene caso acusativo (por ejemplo, *Juan da un libro*) y en la definición de categoría para transitivos que toman un complemento de frase preposicional tiene caso preposicional *p-case* (por ejemplo *Juan se lo dio a María*).

DITRANS–CAT OBL–TRANS–CAT

[RES: strict–trans–cat [RES: strict–trans–cat

ACT: [np–sign ACT: [np–sign

CAT: np[acc]]] CAT: np[p–case]]]

Los restantes tipos de categorías están organizados en *comp-cat* para verbos que toman un complemento oracional y en *xcomp-cat* para verbos que toman un complemento predicativo, los *xcomp-cat* además se dividen de acuerdo a si el control está involucrado o no.

Los *signos de los verbos* se definen enlazando signos activos en la estructura de categorías a las ranuras de argumento en estructuras de argumentos de predicados, es decir, los enlaces se hacen a través de las estructuras semánticas y de categorías. Estos enlaces se realizan mediante enlaces reentrantes, por ejemplo, con la marca [1] en la estructura que se muestra para verbos intransitivos estrictos.

[strict–intrans–sign

CAT: ACT: [np–sign

SEM: [1] <e1>p–agt(e1, x)]

SEM: [strict–intrans–sem

<e1> and (<e1> pred (e1), [1])]]

Solamente consideran patrones para verbos que tienen un máximo de 3 argumentos por lo que solamente necesitan dos patrones adicionales de enlace general.

[dos–argumentos–verbo–signo

CAT: [RES: [RES: sent

ACT: [sign

SEM: [1]]]

ACT: [sign

SEM: [2]]]

SEM: <e1> and (and (pred(e1),[1]),[2])]

[tres–argumentos–verbo–signo

CAT: [RES: [RES: [RES: sent

ACT: [sign

SEM: [0]]]

ACT: [sign

SEM: [1]]]

ACT: [sign

SEM: [2]]]

SEM: <e1> and (and (and (pred (e1),[0]),[1]),[2])]

Finalmente, a continuación se presentan las estructuras completas de dos-argumentos-verbo-signo y de tres-argumentos-verbo-signo. En los primeros se consideran el tipo transitivo estricto y para sujetos de verbos *equi* que toman un complemento de verbo en infinitivo. En los segundos se consideran los ditransitivos y los transitivos que toman un objeto oblicuo.

DOS–ARGUMENTOS–VERBO–SIGNO

STRICT–TRANS–SIGNO
[CAT: strict–trans–cat
SEM: strict–trans–sem]

SUJ–EQUI–INTRANS–GVINF–SIGNO
[CAT: intrans–vpinf–control–cat
SEM:p–agt–subj–intrans–xcomp/comp–sem]

TRES–ARGUMENTOS–VERBO–SIGNO

DITRANS–SIGNO
[CAT: ditrans–cat
SEM: obl–trans/ditrans–sem]

OBL–TRANS–SIGN
CAT: [RES: strict–intrans–cat
ACT: [np–sign
CAT: np[p–case]]]
SEM: intrans–obl–sem]

Los argumentos subcategorizados se posicionan en la estructura de categorías de predicados de acuerdo a la jerarquía oblicua. Por ejemplo, el argumento del sentido "meta" de ditransitivos y de transitivos que subcategorizan un grupo preposicional (DITRANS–SIGNO y OBL–TRANS–SIGN) es el signo extremo en la estructura de categorías, aunque solamente en los ditransitivos le precede el objeto "tema". La diferencia en el orden de palabras se maneja sintácticamente [Sanfilippo, 93].

Este formalismo emplea categorías de control para describir la estructura sintáctica de los verbos *equi* y de ascensión. Crea un modelo donde la marca de reentrancia dice que el signo activo del complemento (por ejemplo un complemento sujeto) se controla por el signo activo inmediatamente precedente. Todas las categorías de control heredan este modelo. El control se expresa mediante entidades que se igualan y que parcialmente describen la semántica de los signos activos. El argumento controlador puede ser el sujeto o el objeto según si el verbo es transitivo o intransitivo. La transitividad está determinada por la presencia de un signo-np acusativo activo. Las categorías reales de control se construyen agregando más especializaciones a las descripciones de control básicas.

En cuanto al trato del sujeto de verbos de extraposición, la CG emplea adicionalmente una entidad sin contenido, *dummy*, para la caracterización semántica de grupos nominales pleonásticos.

Subcategorización en HPSG

En la HPSG, existe una característica especial para la información de la subcategorización de los signos, la característica sintáctica local SUBCAT. En la característica SUBCAT se codifican las diversas dependencias entre un núcleo-*h* y sus complementos. Es de notar que a diferencia de otros formalismos, en la HPSG se incluyen los sujetos como especificadores.

SUBCAT tiene como valor una lista de synsems (parcialmente especificados). Como se mencionó en la sección 1.2-HPSG, los synsems tienen como valor *local* a CATEGORY y a CONTENT. El atributo CATEGORY de un signo contiene información de su POS, requerimientos de subcategorización y marcadores posibles. El atributo CONTENT provee información de su estructura de argumentos. Así que los signos léxicos pueden ejercer restricciones en la selección y manejo de la categoría tanto como en la asignación de papel y caso.

El Principio de Subcategorización en la HPSG, que es un principio de la gramática universal, maneja el flujo (ascendente en la estructura sintáctica) de la información de subcategorización de las trayectorias de proyección. Este principio se expresa en términos de un valor en forma de lista:

DAUGHTERS | HEAD-DAUGHTER | SYNSEM | LOCAL | CATEGORY | SUBCAT, esta lista se obtiene a su vez, de la concatenación de los valores lista de SYNSEM y de DAUGHTERS (ver sección 1.2-HPSG).

El Principio de subcategorización establece, de forma general, que el valor SUBCAT de una frase es el valor SUBCAT del núcleo-*h* del lexema menos las especificaciones ya satisfechas por algún constituyente en la frase. La versión más reciente de HPSG [Sag & Wasow, 99], separa en dos características, SUJ y COMPLS, la característica inicial SUBCAT [Pollard & Sag, 87, 94] para separar el sujeto de los complementos restantes.

En la HPSG la subcategorización se basa en la definición de la estructura de argumentos y cómo se relacionan los roles con los objetos sintácticos (sujeto, objeto, etc.), en la jerarquía de esos objetos sintácticos, en la selección diferente de las categorías de los argumentos, y en las características morfosintácticas de esas categorías. En la HPSG, la asignación de roles es la conexión entre los constituyentes de una expresión y los constituyentes que están presentes en la situación descrita. Por ejemplo, la entrada léxica para un verbo ditransitivo como *dar* asigna papeles semánticos a sus dependientes subcategorizados.

$$\begin{bmatrix} \text{phon} & \textit{dar} & \\ \text{synsem}\,|\,\text{loc}\,|\,\text{cat} & \begin{bmatrix} \text{subcat} & \langle \text{GN}\,\boxed{1}, \text{GN}\,\boxed{2}, \text{"}a\text{"}\ \text{GN}\,\boxed{3} \rangle \\ \text{content} & \begin{bmatrix} \text{reln} & \text{dar} \\ \text{donador} & \boxed{1} \\ \text{dado} & \boxed{2} \\ \text{receptor} & \boxed{3} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

En la lista SUBCAT se numeran las variables asociadas con los objetos sintácticos, éstos unifican con las variables correspondientes de los roles en la descripción CONTENT. La jerarquía de objetos sintácticos se muestra en la lista SUBCAT, donde el sujeto es el primer elemento, el primer objeto es el segundo elemento, y el tercer elemento es el segundo objeto, como en la frase *Juan da un libro a María*. Cada uno unifica con su correspondiente papel, el sujeto unifica con el donador, el primer objeto unifica con el objeto dado, y el segundo objeto unifica con el receptor. Notar en este ejemplo que la posición de los constituyentes en SUBCAT es primordial para identificar cada uno con su rol semántico.

Como se observa del ejemplo anterior, la concepción jerárquica de los objetos sintácticos es esencial. A excepción del sujeto, que tiene su propia lista de características, los otros objetos sintácticos se definen en términos del orden de la jerarquía, que corresponde a la noción gramatical tradicional de sesgadura de objetos sintácticos, con elementos más oblicuos que ocurren más a la izquierda. Los razonamientos para la teoría jerárquica de objetos sintácticos se basa en cuatro clases diferentes de generalizaciones lingüísticas:

- En el orden de constituyentes. En muchos, pero no en todos los lenguajes, el orden superficial de constituyentes y sus objetos sintácticos parecen estar sujetas a restricciones mutuas. Como en el inglés, notar que en el ejemplo anterior el sujeto y los dos complementos se describen igual en SUBCAT, con grupos nominales y solamente el orden estricto permite identificar cada uno de ellos.

- Que involucran la teoría de control. Los complementos controlados encuentran su controlador en un argumento simultáneo menos oblicuo.

- Sobre el ligamento de pronombres y reflexivos. Las relaciones comando-o (de oblicuo, para establecer la teoría de ligamento en la HPSG) se expresan en términos de jerarquía oblicua.

- Sobre el funcionamiento de reglas léxicas. Por ejemplo, la conversión a pasiva puede promover un último o un penúltimo grupo nominal a una posición de sujeto.

En la HPSG se consideró el hecho de que las dependencias léxicas inciden de

manera crucial en la selección de categoría. Existen restricciones de subcategorización que no pueden reducirse a distinciones semánticas o funcionales. En los ejemplos siguientes, se muestran verbos cuyos sentidos están muy cercanos, pero imponen restricciones específicas diferentes sobre la categoría sintáctica de sus argumentos.

Rosalba confía en Rodolfo/ Rosalba se confía de Rodolfo*

Rosalba se fía de Rodolfo / Rosalba se fía en Rodolfo*

Los verbos de *tener confianza* como *confiar* y *fiarse*, tienen estructuras de argumento similares pero muestran una selección diferente de preposición. El autor muestra como ejemplos los verbos *trust* y *rely* que tienen estructuras de argumento similares pero muestran una selección diferente de categoría, el primero a un GN y el segundo a un GP. Puesto que la selección de categoría y de preposición introductora se realizan en la lista de especificaciones SUBCAT, la descripción SUBCAT será diferente en cada caso dentro de MAJ (núcleo–*h* MAJOR). Para el verbo *trust* se indica un grupo nominal:

trust: SUBCAT <... SYNSEM|LOC|CAT|MAJ GN>

En el caso del verbo *rely*, y de los verbos españoles *confiar* y *fiarse*, SUBCAT no solamente especifica la categoría de su complemento como preposicional sino que también exige la preposición específica, que para *confiar* es *en*:

confiar: SUBCAT <... SYNSEM|LOC|CAT [MAJ P, PFORM *en*]>

La subcategorización se basa también en ciertas características morfosintácticas, como la forma verbal, el caso, etc. Por ejemplo, algunos verbos ingleses como *make* y *force* seleccionan diferentes formas verbales, finita e infinitiva.

*Pat made Kim throw up. /*Pat made Kim to throw up.*

*Pat forced Kim to throw up. /*Pat forced Kim throw up.*

Esta realización se define también en COMPLS indicando la forma de inflexión requerida, mediante la característica VFORM, ver Figura 12 . La descripción del verbo *force*, difiere de la anterior en que en lugar de tener VP[*base*] tiene VP[*inf*]. En español, las construcciones no son tan directas, se emplean otras palabras introductoras como preposiciones y conjunciones. Por ejemplo: *Rosalba obligó a Arturo a estudiar* y *Rosalba logró que Arturo estudiara*.

Otra característica del núcleo-*h* como CASE se emplea para lograr una definición similar en lenguajes con inflexiones de caso, donde algunos verbos semánticamente próximos pueden requerir objetos en casos diferentes.

El Principio de Característica del núcleo-*h*, que filtra las características del núcleo-*h* de un nodo hija al nodo madre, establece que siempre que una forma léxica selecciona un complemento de frase especificado como SYN | LOC | HEAD | CASE *ACC* o como SYN | LOC | HEAD | CASE *NOM*, el núcleo-*h* léxico de ese complemento se especifica de la misma manera. Una situación análoga es el manejo

$$\begin{bmatrix} \text{phon} & \textit{made} & \\ \text{synsem}\,|\,\text{loc}\,|\,\text{cat} & \begin{bmatrix} \text{head} & \begin{bmatrix} \text{maj} & \textit{v} \\ \text{vform} & \textit{finito} \end{bmatrix} \\ \text{valence} & \begin{bmatrix} \text{subj}\langle\ \text{SN}\ \rangle \\ \text{compls}\langle\ \text{VP}[\textit{base}], \text{SN}\ \rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figura 12. Descripción del verbo *make*

de la preposición particular que rige una frase preposicional en lenguajes que carecen de inflexión de caso.

Otro punto importante considerado en la subcategorización es el manejo de preposiciones. HPSG enfatiza el hecho de que el empleo de preposiciones particulares no es predecible semánticamente. Por lo que diferentes verbos que requieren complementos realizados con frases preposicionales requieren valores diferentes para la característica del núcleo-*h* PFORM en ese complemento. Por ejemplo, los verbos *destinar, emplear* y *usar* asignan roles correspondientes a complementos introducidos con diferentes preposiciones.

El director destinó un millón de pesos a la biblioteca.

El director empleó un millón de pesos en la biblioteca.

El director usó un millón de pesos para la biblioteca.

Por último, en la HPSG se realiza un trabajo importante para describir los verbos de *control* y de ascensión. Estos verbos tienen como complemento un grupo verbal y el sujeto de este grupo está identificado con un argumento del verbo. La diferencia entre estas construcciones se describe en las entradas léxicas.

- En los verbos *equi* todos los dependientes subcategorizados tienen asignado un rol semántico. Por ejemplo, un verbo *equi* como *try* subcategoriza un sujeto tipo grupo nominal y un complemento tipo grupo verbal.

- En los verbos de ascensión un dependiente subcategorizado no tiene asignado un rol semántico. La identificación de dependiente no se hace compartiendo la estructura de índices sino compartiendo la estructura del synsem completo del dependiente.

Por ejemplo, el verbo *intentar* asigna el rol de “quien intenta” al sujeto, mediante el índice referencial correspondiente, y el valor CONTENT de su complemento VP al argumento psoa (*parameterised state of affairs*). El índice del

sujeto también está en la estructura compartida con el sujeto del complemento de tipo VP, en la lista SUBCAT.

$$\left[\begin{array}{ll} \text{phon} & \textit{intentar} \\ \text{synsem} \mid \text{loc} \mid \text{cat} & \left[\begin{array}{l} \text{subcat} \left\langle \text{GN } \boxed{1}, \text{VP}\left[\text{subcat} \left\langle \text{GN } \boxed{1} \right\rangle\right]\text{:}\boxed{2} \right\rangle \\ \text{content} \left[\begin{array}{l} \text{reln } \textit{intentar} \\ \text{intentante } \boxed{1} \\ \text{psoa} - \text{arg } \boxed{2} \end{array}\right] \end{array}\right] \end{array}\right]$$

Una frase como *Juan intenta correr* tendría la siguiente descripción, donde el rol del sujeto del verbo en infinitivo se indica en “psoa” del verbo *intentar*:

$$\left[\begin{array}{ll} \text{phon} & \textit{intentar} \\ \text{synsem} \mid \text{loc} \mid \text{cat} & \left[\begin{array}{l} \text{subcat} \left\langle \text{GN } \boxed{1}, \text{VP}\left[\text{subcat} \left\langle \text{GN } \boxed{1} \right\rangle\right]\text{:}\boxed{2} \right\rangle \\ \text{content} \left[\begin{array}{l} \text{reln } \textit{intentar} \\ \text{intentante } \boxed{1} \\ \text{psoa} - \text{arg } \boxed{2} \left[\begin{array}{l} \text{reln } \textit{correr} \\ \text{corredor } \boxed{1} \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right]$$

En los verbos de ascensión, que aceptan todo tipo de sujeto se omite la categoría y no se comparte la estructura del sujeto por lo que no está asignado a un papel en la matriz *psoa*. Entonces, la lista SUBCAT especifica que el synsem completo de su sujeto es la estructura compartida con el synsem de su complemento subcategorizado tipo grupo verbal.

Valencias Sintácticas en DUG

En los árboles de dependencias cada nodo representa un segmento elemental (una categoría terminal) por lo que los nodos están típicamente marcados por lexemas. En la DUG, donde no se consideran etiquetas en los enlaces, se prefiere una representación en línea en lugar del árbol, así que por ejemplo la frase *El niño pequeño atrapó una lagartija* puede representarse en la siguiente forma:

[*atrapar* [*niño* [*el*] [*pequeño*]] [*lagartija* [*una*]]]

Esta es una forma equivalente a una estructura jerárquica. En este tipo de representación, DUG a diferencia de otras gramáticas dependencias incluye las categorías de POS a las marcas de los nodos, por ejemplo:

[V *atrapar* [N *niño* [Det *el*] [ADJ *pequeño*]] [N *lagartija* [Det *una*]]]

Donde Det significa determinante y ADJ adjetivo. En la misma forma y combinando categorías funcionales y morfosintácticas DUG introduce ambas categorías en la representación, por ejemplo:

[PRED *atrapar* V

[SUJ *niño* N [DET *el* Det] [ATR *pequeño* Adj]]

[OBJD *lagartija* N [DET *una* Det]]]

Donde PRED es predicado, ATR es atributo, DET es determinante y OBJD es objeto directo. El orden de palabras, que es importante para el inglés, se describe en DUG mediante un marcaje adicional. Por el símbolo '<' para denotar *a la izquierda* del núcleo-*h* y '>' para denotar *a la derecha* del núcleo-*h*, de esta forma se describe que el sujeto está a la izquierda del verbo y el objeto directo a la derecha:

[PRED *atrapar* V

[< SUJ *niño* N [DET *el* Det] [ATR *pequeño* Adj]]

[> OBJD *lagartija* N [DET *una* Det]]]

En la DUG se combina la noción de estructura de frase con la de dependencias ya que considera las dependencias como una relación de palabra a complemento, en lugar de una relación de palabra a palabra, donde un complemento puede consistir de muchas palabras. Es por esta razón que incluye las categorías gramaticales. Por ejemplo, el constituyente *el niño pequeño* es el sujeto del verbo *atrapar* en los ejemplos anteriores.

La DUG considera que internamente, cualquier frase se estructura de acuerdo a las relaciones de palabra a complemento y que se representa como tal. Por lo que aunque todos los nodos hoja en un árbol de dependencias corresponden a elementos terminales, en la DUG los nodos interiores pueden ser no-terminales. Sin embargo, una relación de dependencias solamente existe entre una palabra en el nodo dominador y las frases enteras representadas por el subárbol dependiente. Los nodos en el árbol de dependencias tienen las siguientes características:

- Hay un orden de secuencia entre los dependientes del mismo núcleo-*h*, igual que en la GPSG.
- Los nodos en el árbol representan unidades función-lexema-forma (función sintagmática, significado léxico, características morfosintácticas)
- Los nodos tienen etiquetas múltiples, por ejemplo, numero[singular], género[masculino], no pueden ser estructuras.
- Cada nodo hoja en el árbol corresponde a un terminal y cada subárbol corresponde a un no-terminal.

Un ejemplo se presenta con la frase *Arturo presenció la riña estudiantil*. con la siguiente representación del analizador sintáctico, donde omitimos la posición de cada palabra de la frase:

(ILLOC: *postulado'*: sign

(< PROPOS: *presenciar pasado'*: verbo forma[finita] persona[él, 3, sing]

(<SUBJECT: *Arturo*: sustantivo persona[él, NP, sing] determinado [+, NP])

(>DIR_OBJ1: *riña*: sustantivo persona[3, sing] determinado[+,C]

(DETER: *definido'*: artículo determinado[+,D] (referencia[definido,sing])

(<ATTR_NOM: *estudiantil*: adjetivo determinado[-]))));

En la representación anterior, sin entrar en detalles, se muestra un árbol de dependencias con seis nodos, un nodo para cada palabra de la frase más el nodo raíz que corresponde a la oración. El punto origina el *postulado'* inicial, por lo que el nodo raíz corresponde a la oración, como en el enfoque de constituyentes. Cada nodo lleva tres tipos de información:

- Una función sintáctica, como sujeto SUBJECT, primer objeto DIR_OBJ1, determinante DETER, etc.
- Un lexema, como: *presenciar pasado', Arturo, riña, definido', estudiantil*)
- Un conjunto de características morfosintácticas; la primera característica es la categoría gramatical, como artículo, adjetivo, etc.

El árbol de dependencias se construye a partir de la información contenida en tres diccionarios: un diccionario morfosintáctico, un conjunto de patrones de valencias y un diccionario de valencias.

El *diccionario morfosintáctico* relaciona cada forma de palabra a un lexema y a una categoría morfosintáctica compleja.

Los *patrones de valencia* contienen los fragmentos de un árbol de dependencia, generalmente correspondientes a un rector y un dependiente. Describen relaciones sintagmáticas específicas, entre el nodo del núcleo-*h* y su nodo dependiente (denominado ranura[23]), por ejemplo la relación entre un verbo y su sujeto. En estos patrones se describe la capacidad de combinación de las palabras, en las ranuras se acomodan los elementos de su contexto. Cada patrón caracteriza la forma morfosintáctica del núcleo-*h*, la función sintáctica del dependiente y la forma morfosintáctica del dependiente. También las selecciones léxicas pueden especificarse en una ranura cuando se requiere.

El *diccionario de valencias* consiste de referencias. Una referencia asigna un

[23] *Slot*, en inglés.

patrón o un conjunto de patrones al elemento léxico, de esta forma se implementa la subcategorización, que describe la capacidad de combinación del elemento. Existen tres tipos de referencias de acuerdo a las posibles funciones de los patrones: complementos, adjuntos y conjunciones.

Para el ejemplo anterior, se tienen los siguientes patrones:

(ILLOC: +postulado: signo

(<PROPOS :=: verbo forma[finita] s_type[postulado]));

(*:+subject: verbo forma[finita, indicativo] s_type[postulado, relativo]

(<SUBJECT:=: sustantivo persona[NP] determinado[+]));

(*:+dir_obj1:verbo obj_number[singular] modo[activo]

(>DIR_OBJ1:=: sustantivo persona[1, 2, 3, sing., plural] determinado[+]));

(*: %dete_count_any: sustantivo count[+]

(<DETER: determinante determinado[D]));

(*: %attr_nominal: adjetivo

(<ATTR_NOM: adjetivo determinado[-]));

Las referencias que se emplearon para enlazar los elementos léxicos en la frase del ejemplo con los patrones anteriores son las siguientes:

(:COMPLEMENTS (*:*postulado'*: signo) (: +propos));

(:COMPLEMENTS (*:*presenciar*: verbo) (&(: +subject) (:+dir_obj1)));

(:ADJUNCT (*:definido: determinante) (: %dete_count_any));

(:ADJUNCT (*: *estudiantil*: adjetivo) (: %attr_nominal));

En la DUG se separan completamente los complementos y los adjuntos. Los complementos son dependientes de un elemento léxico y son requeridos por la semántica combinatoria inherente de la palabra. Los adjuntos son circunstanciales, por ejemplo los adverbios. Mientras que un término esta incompleto hasta que ha encontrado sus complementos, los adjuntos pueden agregarse al conjunto de dependientes de un término en una forma relativamente arbitraria. Mientras los complementos se especifican en el diccionario bajo el lema del término rector, es decir, en forma descendente, los patrones adjuntos se especifican en la entrada léxica de la palabra adjunta, definiendo el potencial del enlace del elemento léxico como un dependiente, es decir, en una forma ascendente.

Para describir las alternaciones sintácticas del verbo se aceptan más de un patrón con el mismo nombre. Por ejemplo, entre los patrones de sujeto están los siguientes, que describen los sujetos en oraciones interrogativas:

(*:+subject: verbo inicial[+] forma[finita, indicativo] s_type[pregunta]

(>SUBJECT:=: sustantivo persona[NP] determinado[+]));

(*:+subject: verbo forma[finita, indicativo] s_type[interrogativa, relativa]
(<SUBJECT:=: pronombre pro_form[interrogativa, relativa[C] persona[C] número[sing] caso[de sujeto]));

El primer patrón del sujeto describe el sujeto de *¿Presenció Arturo la riña estudiantil?* y el segundo patrón considera la frase *¿Quién presenció la riña estudiantil?*. Ambos patrones están ya cubiertos por la referencia para *presenciar* en las referencias anteriores.

En la DUG, las estructuras de control y extraposición se manejan por asignación de patrones específicos a los verbos que dan origen a estas estructuras. DUG describe la estructura de argumento como un nivel de descripción sintáctica. No hay un orden de roles participantes, por lo que el sujeto se considera como un argumento más del verbo.

Valencias Sintácticas en la MTT

En los árboles de dependencias de la MTT [Mel'cuk, 79], los arcos entre los nodos están etiquetados con *relaciones sintácticas de superficie*. Estas relaciones son dependientes del lenguaje y describen construcciones sintácticas particulares de lenguajes específicos. Entre estas relaciones, existen unas cuantas donde el dependiente se denomina actuante sintáctico de superficie.

Los actuantes sintácticos de superficie de un verbo representan lo que en otros formalismos se conocen como los objetos sintácticos, es decir, su sujeto, sus objetos y sus complementos pero únicamente relacionados al sentido inherente del lexema. Los actuantes corresponderían a los "complementos" de la DUG ya que contrastan con los circunstanciales (o adjuntos en la DUG). La línea divisoria entre ellos se marca de acuerdo a diversos criterios que se expondrán en otras secciones.

La construcción de la estructura sintáctica de superficie se realiza mediante tres tipos de reglas: 1) las reglas que transforman una relación sintáctica profunda en una relación sintáctica de superficie y viceversa, 2) las reglas que transforman una relación sintáctica de superficie en un nodo de la sintaxis profunda y viceversa, y 3) las reglas que transforman una relación sintáctica profunda en un nodo de la sintaxis de superficie y viceversa. En [Mel'cuk, 88] se presentan estas reglas con ejemplos para el inglés y el ruso.

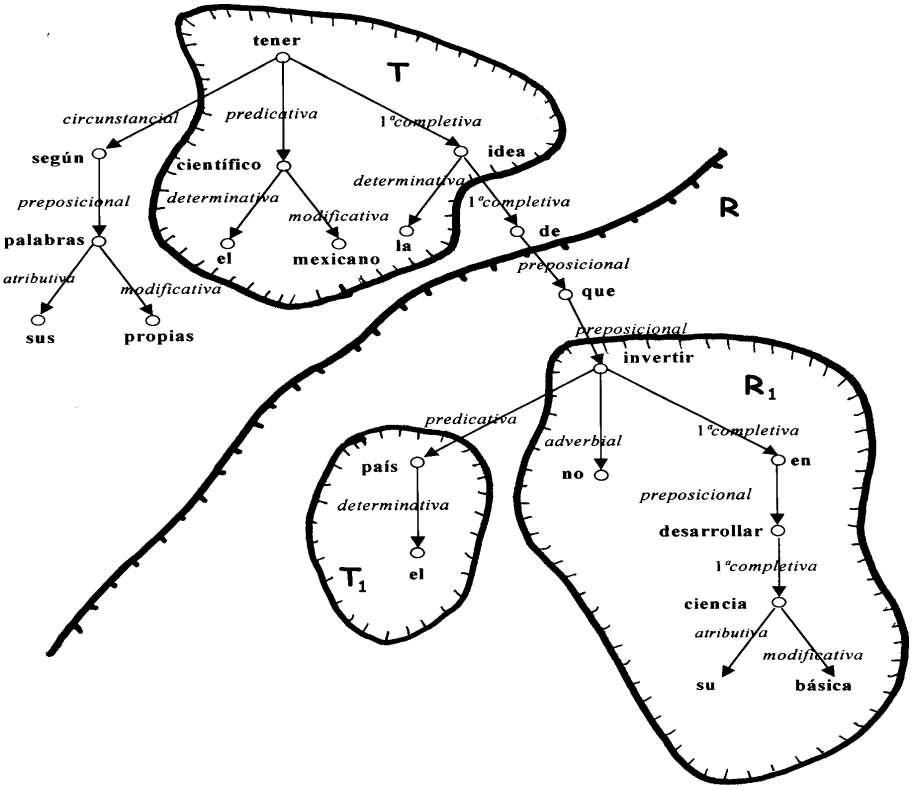


Figura 13. Ejemplo de una representación sintáctica superficial**.**

En el primer tipo se expresan las relaciones sintácticas profundas mediante una relación sintáctica de superficie, por ejemplo, las predicativas, posesivas, modificativas, cuantitativas, etc. En el segundo tipo un lexema profundo ficticio se expresa mediante una relación sintáctica de superficie, por ejemplo, la aproximativa-cuantitativa en el ruso. En el tercer tipo, una relación sintáctica profunda se expresa mediante una palabra función, por ejemplo, las preposicionales.

En la Figura 13 presentamos el diagrama de la representación sintáctica de superficie para la frase *Según sus propias palabras, el científico mexicano tiene la idea de que el país no invierte en desarrollar su ciencia básica.*

En la MTT, las valencias sintácticas de los verbos, principalmente, de los sustantivos, y de los adjetivos se describen conforme a lo que se denomina Zona Sintáctica [Steele, 90], con la ayuda de una tabla de Patrones de Manejo sintáctico (PM). La descripción en esta zona corresponde al nivel de la representación sintáctica de superficie de la MTT, a la estructura sintáctica de superficie.

Existen otras tres estructuras en este nivel (la estructura comunicativa, la

estructura anafórica y la estructura prosódica) que están más relacionadas con la representación sintáctica profunda. En la Figura 13 se observan la estructura comunicativa, el tema y el rema. Con línea punteada se marcan las referencias concurrentes, correspondientes a la estructura anafórica; en este caso la prosodia se considera neutral. Las líneas completas marcan la estructura sintáctica de superficie.

En la tabla de PM de la zona sintáctica, que expresa la diátesis, se presenta la siguiente información:

- Correspondencia entre las valencias semánticas y sintácticas de la palabra encabezado.
- Todas las formas en que se realizan las valencias sintácticas.
- La indicación de obligatoriedad de la presencia de cada actuante, si es necesario.

Así que cada PM es una colección completa de descripciones de todos los posibles objetos de una palabra específica (verbo, sustantivo o adjetivo), sin considerar su orden en la oración.

Después de la tabla de PM, en la zona sintáctica, se presentan dos secciones: restricciones y ejemplos. Las restricciones consideradas en los PM son de varios tipos: semánticas, sintácticas o morfológicas; entre estas restricciones también se considera la compatibilidad entre valencias sintácticas. La sección de ejemplos cubre todas las posibilidades: ejemplos para cada actuante, ejemplos de todas las posibles combinaciones de actuantes y finalmente los ejemplos de combinaciones imposibles o indeseables, es decir, los órdenes permitidos y prohibidos de estas diferentes palabras manejadas.

La parte principal de la tabla de PM es la lista de valencias sintácticas de la palabra encabezado. Se listan de una manera arbitraria pero se prefiere el orden de incremento en la oblicuidad: sujeto, objeto directo, objeto indirecto, etc. Cada encabezado usualmente impone cierto orden; por ejemplo, una entidad activa (sujeto) toma el primer lugar, después el objeto principal de la acción, después otro complemento (si existe), etc. También la forma de expresión del significado de la palabra encabezado influye en el orden. Esta expresión precede cada PM.

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

A continuación presentamos dos descripciones para el vocablo *enseñar*, como una entrada del diccionario explicativo combinatorial, ejemplos para el inglés se presentan en [Steele, 90] y para el francés en [Mel'cuk *et al*, 84, 88]. Para el vocablo *teach*, [Steele, 90] presenta ocho descripciones.

1) X, teniendo conocimiento y habilidades en Y, causa que Z de forma intencional y metódica aprenda Y1 *[El profesor enseña historia a sus estudiantes]*

2) X contiene un postulado Y1, el cuál es parte de una teoría Y2, expuesta en X para la información de Z *[El Capital nos enseña que podemos organizarnos socialmente]*

Cada una de estas descripciones presenta un sentido atribuido al lexema. Cada sentido tiene una forma de realizar sintácticamente sus valencias. La descripción de la zona sintáctica del sentido 1) se presenta en el ejemplo anterior, terminando con un ejemplo.

De lo anterior se desprende que las descripciones propuestas están dirigidas al ser humano. Las entradas del diccionario combinatorio son exhaustivas, indicando todos los posibles sentidos atribuidos al vocablo y con las realizaciones sintácticas de las valencias. Las posibles combinaciones se muestran con ejemplos muy completos.

1 X teaches Y to Z = X, having knowledge of, or skills in, Y, causes Z intentionally and methodically to learn Y1

1 = X	2 = Y	3 = Z
1. N	1. N 2. *a* V_{inf} Obligatory	1. *a* N 2. Pron

$C_1 + C_2$: *El profesor enseñó la teoría de la relatividad; La algoritmia enseña a mecanizar la intuición.*

$C_1 + C_2 + C_3$: *La maestra le enseñó a tocar el piano; La pianista enseñó las escalas a los principiantes; El Dr. Mel'cuk nos enseñó los fundamentos de su teoría; El delegado enseñó al personal a levantar las actas administrativas.*

Ejemplos

El tlamatani, en su profesión de maestro, de muchas formas enseñaba el camino que había que seguirse, con su sabiduría iluminaba lo que está sobre la tierra. Enseñaba a sus discípulos a conocerse a sí mismos; con una metáfora se nos dice que, con tal propósito , "les ponía un espejo delante de sus rostros".

DEFINICIONES LEXICOGRÁFICAS

Las palabras de cada lenguaje natural se dividen en autónomas y auxiliares. Existen unos diccionarios especiales que explican el sentido de cada palabra autónoma. Se llaman diccionarios de la lengua, o de explicaciones y se dirigen a seres humanos. El rasgo muy importante de la MTM es que el diccionario computacional se propone como la estructura que contiene las explicaciones (definiciones lexicográficas) para palabras autónomas, y estas definiciones sirven como el medio para las transformaciones en el nivel semántico, así como para establecer las correspondencias entre las valencias semánticas y sintácticas. En la forma inicial, las definiciones se representan como una oración o un conjunto de oraciones en lenguaje natural. Los rasgos muy importantes de las definiciones son:

- Las palabras usadas se libran de toda ambigüedad, es decir son de un solo significado. Puesto que las palabras comunes de cada lenguaje tienen frecuentemente homónimos, se hace la selección y las marcas especiales.
- El sentido de muchas palabras, especialmente de verbos y sustantivos verbales, no puede definirse sin mencionar unas entidades las cuales hay que precisar en la situación específica. Estas entidades sirven como los papeles en las acciones que son reflejadas por los verbos correspondientes. Son justamente las valencias semánticas del verbo. En las definiciones lexicográficas, las valencias se representan como variables en las formulas algebraicas por letras X, Y, Z, W....
- Debemos explicar el sentido de la palabra por sentidos de otras palabras que son más “simples” que la palabra bajo definición. No tenemos lugar para explicar cual es esta simplicidad, sólo hacemos notar que el conjunto de todas las definiciones no debe contener círculos viciosos y conducir a unos sentidos elementales.

EJEMPLOS DE DEFINICIONES

Las definiciones de clasificación son bastante comunes en los diccionarios de explicación orientados a los seres humanos. En primer lugar dan una noción de cual es el género semántico (= superclase) para la noción bajo definición y además añaden las propiedades específicas de esta especie (= subclase) que le distinguen de otras especies dentro de la misma clase.

Por ejemplo, la definición para *arándano* dice:

arándano *es una baya comible de color azul o negruzco*

Entonces podemos representar esta fórmula de lenguaje natural con la fórmula lógica usando predicados *ES_SUBCLASE*(), *AZUL*(), *NEGRUZCO*() y *COMIBLE* ():

ES_SUBCLASE(arándano, baya) & COMIBLE(arándano) & (AZUL(arándano) ∨ NEGRUSCO(arándano))

A su vez el predicado *COMIBLE* puede expresarse con *COMER*() e *INSALUBRE*() que se consideran más simples:

$COMIBLE(y) \equiv \sim\exists_{persona}\ INSALUBRE\ (COMER\ (persona, y), persona))$

(*Es comible* ***y*** *= No existe persona para la cual es insalubre comer* ***y***)

Las definiciones de unos predicados por otros son también bastante comunes. Si definimos *soltero* en una forma libre como

Soltero es un hombre adulto para quien no existe una mujer con la cual él esté casado

podemos expresar el predicado *SOLTERO*() con los predicados *SEXO*(), *ADULTO*() y *CASADO*():

$SOLTERO(x) \equiv SEXO(x,$ masculino) & $ADULTO(x)$ & $\sim\exists_y$ ($SEXO(y,$ femenino) & $CASADO(x, y)$)

Este es el método de convertir las formulas libres de las definiciones a las fórmulas lógicas correspondientes. Pero el problema de seleccionar las palabras sin homónimos y círculos viciosos en las fórmulas libres es bastante complejo. Al mismo tiempo palabras de lenguajes extranjeros parecen más exentas de homonimia. Es por esto que preferimos las definiciones en inglés para la descripción de sentidos.

Métodos lexicográficos tradicionales de compilación de diccionarios

La lexicografía es una rama de la lingüística aplicada que trata con el diseño y la construcción de bases de datos léxicas (diccionarios, enciclopedias) para el uso práctico de los seres humanos y de sistemas tecnológicos. También trata con su adecuación a cometidos generales o específicos y con el acopio de los recursos teóricos necesarios para alcanzar estos fines.

Los métodos lexicográficos difieren dependiendo de los objetivos y las fuentes de información. Por ejemplo, un diccionario clásico puede tener las siguientes características de representación durante el proceso de desarrollo lexicográfico: 1) un formalismo de estructuras de campos como bases de datos para entradas léxicas, con referencias cruzadas a otros campos, 2) un número de notaciones, para diferentes campos, o para léxico diferente basado en la misma base de datos lexicográfica, y 3) varias implementaciones (como bases de datos). Pero para construir un diccionario clásico en base a un corpus de textos, se requieren varios pasos adicionales [Gibbon, 99]:

1. Adaptación de conjuntos de caracteres, de estructuras de registros, etc. a los requerimientos del marco de trabajo del lexicógrafo.
2. Identificación de las unidades estructurales más pequeñas del texto de entrada, palabras, y resolución de elementos codificados (datos, abreviaturas, etc.)
3. Identificación de las formas de palabra completamente flexionadas que aparecen en el contexto del corpus, que servirá de fuente de información.
4. Especificación de la microestructura: definición de la estructura de los atributos, de la estructura del registro de la base de datos, etc. para los tipos de información léxica que se requiere.
5. Extracción de información:

   (a) análisis estadístico, en sus diferentes variantes (frecuencia de las palabras, frecuencia de pares de palabras, frecuencia de colocaciones, estimación de la probabilidad como información de la microestructura, etc.)

   (b) análisis lingüístico, es decir, lematización (extracción de palabras encabezado), información fonológica, ortográfica, morfológica, sintáctica, semántica y pragmática de microestructura.

En la construcción de diccionarios computacionales, los investigadores hacen énfasis en la distinción de entradas mediante el sentido. Los principios para identificar un sentido en lexicografía según [Meyer *et al*, 90] y [Mel'cuk, 88a], son los siguientes:

1. Si para una unidad léxica sugerida, pueden aplicar dos posibles mapeos a la ontología[24], entonces se deben crear dos unidades léxicas (es decir, crear dos sentidos si se desea tener significados diferentes apuntando a diferentes partes de una jerarquía de tipos).
2. Si hay restricciones elegibles incompatibles para una unidad léxica sugerida, debe haber dos sentidos.
3. Si hay dos conjuntos incompatibles de ocurrencias concurrentes (morfológicos, sintácticos como marcos de subcategorización, o léxicos como colocaciones), se deben crear dos sentidos.
4. Si hay dos posibles lecturas de una palabra, se deben crear dos sentidos.

La creación de entradas en el diccionario ha sido una tarea manual cuyo trabajo requiere expertos. [Mel'cuk, 88a] establece criterios para distinguir sentidos,

[24] La ontología provee un sistema de conceptos (identificación de conceptos y cualquier relación entre ellos). Cada sentido de la palabra se enlaza a algún concepto en la ontología, que se espera sea independiente de lenguajes particulares.

criterios que están dirigidos a los humanos. Para él, un vocablo es el conjunto de todas las unidades léxicas (sentidos) para el cuál las definiciones lexicográficas están ligadas con un puente semántico. Un puente semántico entre dos unidades léxicas es una componente común a sus definiciones, que formalmente expresa un enlace semántico. Una unidad léxica básica de un vocablo es una unidad léxica que tiene un puente semántico con la mayoría de las otras unidades léxicas del vocablo.

Un campo semántico es el conjunto de todas las unidades léxicas que comparten una componente semántica no trivial explícitamente distinguida. Un campo léxico es el conjunto de todos los vocablos cuyas unidades léxicas básicas pertenecen al mismo campo semántico. Aunque Mel'cuk usa un vocablo para agrupar sentidos similares bajo una *superentrada*, cualquier entrada principal puede tener cualquier número de grupos de sentidos bajo ella.

Mel'cuk articula el principio de descomposición donde la definición de una unidad léxica debe contener solamente términos que son semánticamente más simples que la unidad léxica. Más aún, a través de su principio de puente semántico, las definiciones de cualesquiera dos unidades léxicas del mismo vocablo deben enlazarse explícitamente, ya sea por un puente semántico o por una secuencia de puentes semánticos.

Estos principios deben seguirse en la construcción de un diccionario y asegurar su consistencia interna. Más importante aún es que estos principios deben aplicarse para determinar la relación entre una definición y el resto del diccionario, incluyendo otras definiciones de la misma entrada principal. Mel'cuk hace seis observaciones pertinentes para agrupar y ordenar los sentidos de una entrada:

1. El agrupamiento en un vocablo polisémico tiene una motivación semántica, es decir, que todos los lexemas deben compartir al menos un componente semántico importante.
2. La división en grupos de sentidos está basada semánticamente.
3. El ordenamiento se basa en proximidad semántica.
4. El ordenamiento se basa en cuál entrada es semánticamente más simple.
5. Un sentido intransitivo se sitúa antes de un sentido transitivo, de nuevo basado en simplicidad semántica (el transitivo se define en términos del intransitivo).

[Litkowski, 92] considera como principios lexicográficos para organizar un diccionario computacional, los siguientes: las entradas principales y palabras encabezado, el agrupamiento y el orden de sentidos, y por último las seudoentradas. Las entradas principales y palabras encabezado, se refiere a que las unidades léxicas en un diccionario generalmente tienen la intención de asegurar la lexicalización del significado, uniendo grupos y configuraciones de elementos semánticos en unidades léxicas reales y proveyendo información sintáctica y léxica de ocurrencia concurrente.

Pueden existir varias entradas correspondientes a homónimos.

El agrupamiento y el orden de sentidos se refiere a que la creación de sentidos para un diccionario computacional tiene consecuencias importantes para el compromiso del análisis sintáctico que se implemente. En diccionarios para sistemas amplios, mientras más información se tenga en el diccionario, la estructura de una entrada supone mayor importancia, particularmente la manera en la cuál los sentidos se relacionan uno a otro.

Las seudoentradas se refieren a que se codifica otro grupo distinto de entrada léxica para caracterizar generalidades lingüísticas y léxicas. Las seudoentradas codifican solamente abstracciones semánticas o gramaticales, constituyen entradas metalingüísticas en el diccionario. Las seudoentradas varían en importancia con la teoría gramatical.

[Ilson & Mel'cuk, 89] discuten varios problemas léxico-gramaticales: las cuasi-pasivas, las variaciones sintácticas y los complementos objeto y sujeto. Las cuasi pasivas no son posibles en todos los verbos, son lexemas separados de sus formas activas, mientras que las pasivas reales son formas gramaticales del mismo lexema. Por lo que discuten que las pasivas reales no se deben describir como entradas separadas en las entradas propias del diccionario.

La variación sintáctica se refiere a que puede haber dos patrones de manejo que tengan el mismo significado para un solo sentido de un verbo. Por lo que discuten que solamente es necesario un sentido en el diccionario. En los complementos sujeto / objeto, algunos son obligatorios y deben incluirse entre los argumentos de los verbos correspondientes, mientras que otros son opcionales y añadidos libremente. Así que arguyen que el reconocimiento debe tratarse en la gramática y no como resultado de diferentes entradas.

En todos estos casos, cierta información puede situarse en el diccionario. Tal vez la clave para hacer distinciones sea la eficiencia en el procesamiento, por ejemplo, situar información en el diccionario si puede tenerse acceso a ella, y usarse más eficientemente que el retroceso a través de varias trayectorias en un analizador sintáctico. Con el desarrollo de reglas léxicas, reglas derivacionales, y funciones de colocación que pueden situarse en el diccionario mismo, es difícil determinar exactamente dónde abandonar la creación de entradas del diccionario, es decir, en qué momento detener las definiciones lexicográficas.

Revisión de los enfoques diversos para la descripción de valencias sintácticas

En todos los formalismos descritos, las valencias sintácticas involucran tanto la estructura de los distintos argumentos como la función gramatical de cada uno de ellos. El número de argumentos y la descripción de la función gramatical que cada

uno de los formalismos considera difiere, así como el nivel en que se representan.

La estructura de argumentos, es decir, los predicados y los argumentos asociados con los participantes, se define en el nivel sintáctico en la GB, en la GPSG, en la LFG, en la DUG, y en la MTT; en cambio en la HPSG y en la CG forma parte de la representación semántica de predicados.

Los participantes de la acción en todos los formalismos con la excepción de la HPSG, la DUG y la MTT se marcan con roles temáticos que no están motivados totalmente de manera semántica. En la HPSG, la DUG y la MTT se marcan los participantes específicos del significado de cada verbo o palabra de que se trate. Se hace clasificación de roles temáticos en la GB (externos e internos), en la LFG (una jerarquía temática universal) y en la CG (roles prototípicos de Dowty, aumentados). Esta clasificación determina la funcionalidad sintáctica de los participantes.

Por la importancia de la selección semántica en la subcategorización, formalismos como la GB o la LFG que no incluyen un nivel de representación semántica proveen un nivel de descripción lingüística que expresa la estructura semántica de los objetos de los predicados en términos sintácticos.

Mientras en la DUG y en la MTT los objetos sintácticos se expresan léxicamente y se ven como primitivas; en los demás formalismos, los objetos sintácticos se ven como enlaces entre constituyentes seleccionados sintácticamente y los roles semánticos. A excepción de la GB que sitúa esta información en la estructura sintáctica, los demás formalismos la colocan en el diccionario.

La especificación de los objetos sintácticos se hace en la GB como relaciones de predicación y rección; en la LFG la especificación se hace mediante los principios de mapeo léxico, que rige el enlace de roles-θ a las características de las funciones gramaticales primitivas en formas léxicas. En la HPSG y la CG los argumentos se clasifican sintácticamente de acuerdo a la jerarquía oblicua. En la MTT y en la DUG no se define una jerarquía, y aunque se puede emplear el orden en la oblicuidad, existen otros factores a considerar, como el orden de los actuantes en el sentido del lexema.

De entre estos formalismos solamente la LFG y la MTT consideran la estructura de información o comunicativa, en la primera con el foco y tópico, y en la segunda con el tema y el rema. La estructura de información ha sido un problema en el enfoque de constituyentes, porque a menudo las unidades de información no coinciden con las unidades establecidas por la estructura de frase.

1.3 MÉTODOS ESTADÍSTICOS: UNA HERRAMIENTA PARA BÚSQUEDA DE REGULARIDADES

En esta sección presentamos los métodos estadísticos requeridos para reconocer modelos del lenguaje. Estos modelos permiten explicar fenómenos del lenguaje para sistemas computacionales. Por lo que mediante estos métodos estadísticos se detectan regularidades de los lenguajes.

Para emplear métodos estadísticos en el reconocimiento de secuencias de letras y palabras en los lenguajes naturales es necesario primero conocer el concepto de información. [Weaver, 49] estableció que la palabra información en la teoría de comunicación se relaciona no tanto con lo que se dice sino con lo que se puede decir. La información es una medida de la libertad de selección cuando se escoge un mensaje. El concepto de información se aplica no sólo a mensajes individuales, como sería el concepto de sentido, sino a la situación como un todo.

Para aclarar esta situación, un ejemplo es el caso donde el contenido del mensaje depende de echar al aire una moneda. Si el resultado es águila, el mensaje entero consistirá de una palabra, de lo contrario el mensaje consistirá del texto entero de un libro. En este ejemplo, para la teoría de la información lo único importante es que hay dos salidas equiprobables, y no tiene que ver con que el contenido semántico del libro sea mayor que el de una sola palabra. La teoría de la información se interesa en la situación antes de la recepción del símbolo, más que en el símbolo mismo. Por ejemplo, la información es muy baja después de encontrar la letra *q* (en textos en español) puesto que hay una mínima libertad de elección en lo que viene después, porque casi siempre es una *u*.

La cantidad empleada para medir la información es la entropía, exactamente el

término conocido en termodinámica (H)[25]. Si una situación está totalmente organizada, es decir, no está caracterizada por un alto grado de aleatoriedad o elección, la información o entropía es baja.

La unidad básica de información es el bit. El bit se define como la cantidad de información contenida en la elección de uno de dos símbolos equiprobables como 0 ó 1, *si* o *no*. Cada mensaje generado a partir de un alfabeto de *n* símbolos o caracteres puede codificarse en una secuencia binaria. Cada símbolo de un alfabeto de n-símbolos contiene $\log_2(n)$ bits de información, puesto que es el número de dígitos binarios requeridos para transmitir cada símbolo. Por ejemplo para cada uno de los 33 caracteres en el alfabeto para el lenguaje español (a, b, c, …, ñ, …z, á, é, í, ó, ú, ü), se requieren $\log_2(33) = 5.044$ bits.

La entropía está relacionada con la probabilidad. Por ejemplo, cuando se ha empezado a transmitir un mensaje que empieza con las palabras *"se diría"*. La probabilidad de que la siguiente palabra sea *que* es muy alta, mientras que la probabilidad de que la siguiente palabra sea *perico* es muy baja. La entropía es baja en las situaciones donde las probabilidades son muy desiguales y mayor cuando las probabilidades de varias elecciones son iguales. La relación exacta entre entropía y probabilidades inherentes en un sistema está dada por la siguiente fórmula:

$$H = - \left[p_1 \log_2(p_1) + p_2 \log_2(p_2) + \ldots + p_n \log_2(p_n)\right]$$

El signo menos hace que H sea positiva, puesto que los logaritmos de fracciones son negativos. Para calcular la entropía de un lenguaje natural se debe:

- Contar cuantas veces aparece cada letra del alfabeto.
- Encontrar la probabilidad de ocurrencia de cada letra al dividir su frecuencia por el número total de letras en el texto.
- Multiplicar cada probabilidad de letra por su logaritmo base dos.
- Cambiar el signo menos por uno más.

Por ejemplo, la entropía de caracteres de la palabra *lata* se calcula como sigue: *l* ocurre una vez, *a* ocurre dos veces y *t* ocurre una vez. Este minúsculo texto consiste de cuatro letras, la probabilidad de ocurrencia de *l* es 0.25, la de *a* de 0.5 y la de *t* es de 0.25, la probabilidad de todas las otras letras es cero porque no aparecen en el texto. Cuando se multiplica cada probabilidad de letra por su logaritmo de base dos, para *l* se obtiene $0.25 \times \log_2(0.25) = 0.25 \times -2 = -0.5$, para *a* se obtiene $0.5 \times -2 = -1.0$ y para *t* se obtiene -0.5 . Sumando estos valores y cambiando el signo se obtiene el valor de entropía final de 1.5

[Kahn, 66] escribió que el lenguaje con la entropía máxima posible sería aquél

[25] En las ciencias físicas, la entropía asociada con una situación es una medida del grado de aleatoriedad.

sin reglas que lo limitaran. El texto resultante sería completamente aleatorio, teniendo todas las letras la misma frecuencia y cualquier carácter igualmente probable de seguir a cualquier otro carácter.

Sin embargo, las reglas de cualquier lenguaje natural le imponen una estructura y por lo tanto una baja entropía. La fórmula anterior da el grado de entropía de acuerdo a la frecuencia de caracteres solos en el lenguaje, sin tomar en cuenta que la probabilidad de encontrar una letra también depende de la identidad de sus vecinas. Se pueden hacer mejores aproximaciones a la entropía de un lenguaje natural repitiendo el cálculo anterior para cada par de letras (*bigram*) como *ac*, *ad*, etc. después dividiendo entre dos porque la entropía se especifica en una base por letra. Una mejor aproximación aún se produce al realizar el cálculo anterior para cada tres letras (*trigram*) como *adm*, *con*, etc. y después dividiendo entre tres.

El proceso de aproximaciones sucesivas a la entropía puede repetirse incrementando cada vez la longitud del grupo de letras hasta encontrar las secuencias más largas de caracteres (*n-grams*) las cuales ya no tienen una probabilidad válida de ocurrencia en textos. Mientras más pasos se tomen, más precisa será la estimación final de entropía, puesto que cada paso da una aproximación más cercana a la entropía del lenguaje como un todo.

Tomando un alfabeto de 27 letras (26 letras y un carácter espacio), [Shannon, 49] encontró que la entropía del inglés fue de 4.03 bits para una letra, de 3.32 bits por letra en bigrams, y de 3.1 bits por letra en trigrams. La razón de esta disminución es que cada letra influye a la que sigue, es decir, imponen un orden. En base a esto Shannon estableció que cualquiera que hable un lenguaje posee implícitamente un enorme conocimiento de las estadísticas de un lenguaje. Desafortunadamente ese conocimiento es vago e impreciso, por lo que se requieren modelos lingüísticos.

[Edmundson, 63] definió el término *modelo lingüístico* como una representación abstracta de un fenómeno del lenguaje natural. Estos modelos requieren datos cuantitativos así que necesariamente tienen que basarse en corpus. Los modelos lingüísticos pueden ser predictivos o explicativos. Los modelos predictivos expresan la explicación de comportamiento futuro. Los modelos explicativos existen para explicar fenómenos ya observados. Algunos modelos se emplean tanto como modelos predictivos como explicativos, por ejemplo el modelo de Markov.

Un modelo del lenguaje siempre es una aproximación al lenguaje real. Ejemplos de modelos estadísticos del lenguaje son: las predicciones estocásticas de secuencias y los rangos de distribución de frecuencias. El término proceso estocástico fue definido por [Shannon, 49] como un sistema físico, o un modelo matemático de un sistema, que produce una secuencia de símbolos gobernados por un conjunto de probabilidades.

Los modelos lingüísticos basados en estadísticas son necesarios para

considerar la variedad de observaciones lingüísticas y comportamiento cognitivo inherente en la producción de patrones de secuencias de palabras en el lenguaje. Ejemplos de modelos estadísticos del lenguaje son los de [Markov, 16], predicción estocástica de secuencias, el de [Shannon, 49], redundancia del inglés, y el de [Zipf, 35], distribución de rangos de frecuencias.

En esta sección presentamos la distribución de rangos de frecuencias, la predicción estadística de secuencias, y la reestimación.

Distribución de rangos de frecuencias

Entre los modelos predictivos, la ley de Zipf trata de explicar el comportamiento futuro. De acuerdo a la distribución Zipf [Zipf, 49], una variable aleatoria tiene una distribución Zipf si la probabilidad de su función masa esta dada por la siguiente fórmula para algún valor de $\alpha>0$.

$$P\{X = k\} = \frac{C}{k^{a+1}}, \quad k = 1, 2, \ldots$$

Puesto que la sumatoria de las probabilidades anteriores debe ser igual a 1, entonces:

$$C = \left[\sum_{k=1}^{\infty} \left(\frac{1}{k} \right)^{a+1} \right]^{-1}$$

La ley de Zipf dice que para la mayoría de los países, la distribución del tamaño de las ciudades se ajusta impresionantemente a una ley poderosa: el número de ciudades con poblaciones mayores que *S* es proporcional a *1/S*. Suponiendo que, al menos en la última parte, todas las ciudades siguen algún proceso de crecimiento proporcional (esto parece verificarse empíricamente). Esto lleva su distribución, automáticamente, a converger a la ley de Zipf.

De acuerdo a la ley de Zipf, el rango de una palabra en una lista de frecuencias de palabras, ordenada por frecuencias de aparición en forma descendente, está relacionada inversamente a su frecuencia. Se puede predecir la frecuencia de una palabra a partir de su rango usando la fórmula:

$$frecuencia = k \times rango^{-g}, \quad k \text{ y } g \text{ son constantes empíricamente determinadas}$$

La ley de Zipf es una observación empírica de que en muchos dominios, el rango de un elemento dividido por la frecuencia de ocurrencia de ese elemento es constante. Por ejemplo, si las poblaciones de ciudades obedecen la ley Zipf, significaría que si la más populosa tiene una población n, entonces la segunda ciudad más grande tiene n/2 y la tercera n/3, etc. Zipf observó que esta ley se aplica en muchas áreas diversas, incluyendo frecuencias de palabras en textos, escritas en diversos lenguajes. Publicaciones posteriores demostraron que la ley de Zipf es una

consecuencia de asumir que la fuente del lenguaje del cuál se toman los datos de frecuencia es un proceso estocástico simple.

De la fórmula de frecuencia observamos una interdependencia lineal entre frecuencia y rango. Esa fórmula no puede extrapolarse al infinito, puesto que su normalización es imposible. Para los primeros rangos, el cálculo probabilístico directo puede realizarse pero para rangos muy grandes la situación es muy diferente. En cualquier conjunto de frecuencias empíricas, cerca de la mitad de todos los rangos corresponde a los casos de una ocurrencia de los objetos bajo observación. Por lo que objetos con valores grandes de rango no pueden ordenarse apropiadamente, tampoco calcularse con precisión.

La ley de Zipf nos dice que tendremos dificultad delineando cualquier conclusión basada en la observación de la distribución de la mayoría de los elementos del tipo que nos interesa (frase, palabra, etc.). Además de indicarnos que muchos elementos ocurrirán con una frecuencia muy baja, podemos deducir que habrá un gran número de elementos disponibles pero que no ocurren en el corpus de textos. Para nuestra investigación, donde no podemos "alisar" los datos, simplemente nos indica que se requiere incorporar información sobre ellos.

Para tener buenos resultados deberíamos incrementar el tamaño de los datos hasta un límite, mucho mayor que el valor del rango, pero este requisito nos lleva a una labor muy larga de extracción y acumulación de datos, es decir, a una tarea que consume una cantidad impresionante de trabajo.

La distribución de palabras, en varios lenguajes naturales, sigue la ley de Zipf [Baayen, 92], pero la distribución de caracteres concuerda menos. [Shtrikman, 94] muestra que esta diferencia es menos pronunciada en el chino porque muchos caracteres son realmente palabras completas.

Predicción estadística de secuencias aleatorias de palabras

En esta sección presentamos métodos probabilísticos que consideran no un evento aislado sino eventos dependientes, es decir de probabilidades condicionales. La probabilidad condicional de la salida de un evento se basa en la salida de un segundo evento.

MODELO DE MARKOV

Entre los ejemplos que Shannon dio como procesos estocásticos están los lenguajes naturales escritos. Shannon hizo una serie de experimentos para generar un texto, considerando desde el más simple, donde los símbolos son independientes y equiprobables, denominado aproximación de orden cero, hasta el de estructuras de trigram para el inglés. El parecido a un texto usual en inglés aumenta en cada uno de los pasos. En el caso de primer orden, la selección depende solamente de la letra

precedente, nada más. La estructura estadística se puede describir por un conjunto de probabilidades de transición $P_i(j)$, la probabilidad de que la letra *i* sea seguida de la letra *j*. Una forma equivalente de especificar esta estructura es con las probabilidades de bigrams o de la secuencia de dos caracteres P(i, j), la frecuencia relativa del bigram *i*,*j*.

El siguiente paso en complejidad involucra frecuencias trigram. Para esto se requiere un conjunto de frecuencias trigram P(i, j, k) o probabilidades de transición $P_{ij}(k)$. Por ejemplo, los trigram encontrados por [Pratt, 42] para el inglés son: THE, ING, ENT, ION. Arriba de este nivel, se topa uno con la ley de regresos disminuidos y muy grandes, matrices de transición muy poco densas.

Los procesos estocásticos del tipo descrito se conocen como procesos discretos de Markov. La teoría de estos procesos fue desarrollada por [Markov, 16]. En un modelo de Markov, cada estado exitoso depende solamente del estado presente, así que una cadena de Markov es la primera generalización posible, alejada de una secuencia independiente de experimentos. Un proceso complejo de Markov es uno donde la dependencia entre estados se extiende más adelante, a una cadena precedente al estado actual. Por ejemplo, cada estado exitoso puede depender de los dos estados previos. Una fuente de Markov para la cual la selección del estado depende de los *n* estados precedentes da una aproximación de orden (n+1)-iésimo al lenguaje del cual las probabilidades de transición fueron delineadas y se denota como un modelo de Markov de orden *n*-iésimo. Si cada estado exitoso depende de los dos estados previos, tenemos un modelo de Markov de segundo orden, produciendo una aproximación de tercer orden al lenguaje.

Shannon describió los procesos ergódicos de Markov como procesos en los cuales cada secuencia producida de suficiente longitud tiene las mismas propiedades estáticas que las frecuencias de letras y frecuencias de bigrams. En estos modelos cada estado del modelo puede alcanzarse desde cualquier otro estado en un número finito de pasos. El lenguaje natural es un ejemplo de un proceso ergódico de Markov.

Un modelo oculto de Markov (en inglés, *Hidden Markov Model*, HMM) es un proceso doblemente estocástico que consiste de: (a) un proceso estocástico subyacente que no puede observarse, y (b) un proceso estocástico cuyos símbolos de salida pueden observarse, representados por las probabilidades de salida del sistema. Los componentes esenciales de este modelo pueden resumirse en: el conjunto completo de probabilidades de transiciones, el conjunto completo de probabilidades de salida, y su estado inicial. Básicamente, un modelo HMM es un autómata finito en el cuál las transiciones entre estados tienen probabilidades y cuya salida también es probabilística. [Sharman, 89] establece que cuando estos modelos se aplican prácticamente, deben solucionarse tres problemas importantes: evaluación, estimación y entrenamiento.

El problema de evaluación es calcular la probabilidad de que una secuencia de

símbolos observada ocurra como resultado de un modelo dado. En el problema de estimación se observa una secuencia de símbolos producidos por el modelo HMM. La tarea es estimar la secuencia más probable de estados que el modelo realiza para producir esa secuencia de símbolos. Durante el entrenamiento, los parámetros iniciales del modelo se ajustan para maximizar la probabilidad de una secuencia observada de símbolos. Esto permitirá que el modelo prediga secuencias futuras de símbolos.

La solución a la ecuación de la probabilidad de que sea la marca t1 dada la marca previa t0 dada la probabilidad de que la palabra1 tenga la marca t1 tiene al menos dos algoritmos conocidos: Viterbi y backward-forward.

Este modelo ha sido muy empleado en reconocimiento de voz, un tutorial extenso en este tema se encuentra en [Rabiner, 89]. Técnicas estadísticas basadas en HMM están bien establecidas [Holmes, 88] para esa área. En el área de análisis sintáctico, [Collins, 99] usó bigrams, es decir, probabilidades de dependencias entre pares de palabras, como estadísticas para mejorar el análisis sintáctico, emplea el núcleo-*h* del constituyente asociado a otro núcleo-*h* dependiente.

En las llamadas gramáticas de Markov [Charniak, 97] se almacenan las probabilidades que permiten inventar reglas de improviso. Por ejemplo, al inventar reglas de NP se debe conocer la probabilidad de que un NP empiece con un determinante (una probabilidad alta) o con una preposición (una probabilidad baja). Similarmente, al estar creando una frase nominal y con una entrada de determinante se debe saber cual es la probabilidad de que el siguiente constituyente sea un adjetivo (una probabilidad alta) u otro determinante (una probabilidad baja). Sin embrago, estas estadísticas se obtienen de los llamados bancos de árboles (*tree-bank*, en inglés), es decir, corpus analizados y marcados sintácticamente cuya labor manual es intensiva en extremo. También hay que considerar que tienen errores y son limitados.

INFORMACIÓN MUTUA ENTRE PALABRAS DE UNA SECUENCIA

A continuación se describe la llamada *información mutua* en el contexto establecido de la teoría de la información. Considerando *h* e *i* como los eventos que ocurren dentro de secuencias de eventos, en un contexto lingüístico, *h* podría ser una palabra de entrada a un canal ruidoso mientras que *i* es una palabra de salida del canal. *h* e *i* deben ser miembros de la misma secuencia. Por ejemplo, dos palabras que ocurren en una colocación idiomática.

[Sharman, 89] describe cómo la información mutua, denotada $I(h, i)$ muestra qué información se provee del evento *h* por la ocurrencia de *i*. $P(h \mid i)$ es la probabilidad del evento *h* habiendo ocurrido cuando se sabe que el evento *i* ha ocurrido, llamada la *probabilidad a posteriori*; y $P(h)$ es la probabilidad del evento *h* habiendo ocurrido cuando no se sabe si *i* ha ocurrido, llamada la *probabilidad a priori*. La relación entre la probabilidad a posteriori de *h* y la probabilidad a priori de

h es:

$$I(h, i) = \log_2 (P(h \mid i) / P(h))$$

[Church *et al*, 91] establecieron que la información mutua puede emplearse para identificar diferentes fenómenos lingüísticos, desde relaciones semánticas como *doctor – enfermera*, hasta preferencias de ocurrencia simultánea léxico- semántica entre verbos y preposiciones. En las primeras, se encuentra la fuerza de asociación contenida en las palabras, y en las últimas se encuentra la fuerza de asociación entre una palabra contenido y una palabra conexión. Mientras mayor es la información mutua, más genuina es la asociación entre dos palabras.

Este método fue empleado por [Yuret, 98] para encontrar los enlaces entre palabras, sin considerar información gramatical. Aunque obtiene un porcentaje de 60% de precisión entre relaciones de palabras de contenido, una deficiencia es que no se encuentran diferencias entre frases con diferentes preposiciones, por ejemplo *el arquitecto está trabajando en el edificio gubernamental*, y *el arquitecto está trabajando sobre el edificio gubernamental.*

ESTADÍSTICAS BAYESIANAS

Cuando se emplean las estadísticas Bayesianas se trata la probabilidad condicional de una proposición dada una evidencia particular. Es decir, se trata de la creencia en una hipótesis más que su probabilidad absoluta. Este grado de creencia puede cambiar con el surgimiento de nueva evidencia.

La teoría de probabilidad Bayesiana [Krause & Clark, 93] puede definirse usando los axiomas siguientes:

- Primero $p(h|e)$ la probabilidad de una hipótesis dada la evidencia, es una función monótona[26] continua en el rango 0 a 1.
- Segundo, $p(True|e) = 1$, significa que la probabilidad de una hipótesis verdadera es uno.
- Tercero, el axioma $p(h|e) + p(\neg h|e) = 1$ significa que ya sea la hipótesis o su negación será verdadera.
- Cuarto, la igualdad $p(gh|e) = p(h|ge) \times p(g|e)$ da la probabilidad de dos hipótesis que son simultáneamente verdaderas, lo cuál es igual a la probabilidad de la primera hipótesis, dado que la segunda hipótesis es verdadera, multiplicado por la probabilidad de la segunda hipótesis.

Del cuarto axioma se puede actualizar la creencia en una hipótesis en respuesta a la observación de la evidencia. La ecuación $p(h|e) = p(e|h) \times p(h)/p(e)$ significa que la creencia actualizada en una hipótesis *h* observando la evidencia *e* se

[26] Siempre que la evidencia aumenta, la probabilidad de la hipótesis también aumenta, y siempre que la evidencia disminuye la probabilidad de la hipótesis también disminuye.

obtiene la multiplicar la creencia previa en *h*, *p(h)*, por la probabilidad *p(e/h)* de que la evidencia será observada si la hipótesis es verdadera. *p(e/h)* se llama la probabilidad a posteriori, mientras que *p(e)* es la probabilidad a priori de la evidencia. De esta forma, la probabilidad condicional y Bayesiana actualizadas permiten razonar de la evidencia a la hipótesis (abducción) tanto como de la hipótesis a la evidencia (deducción).

Otra consecuencia del cuarto axioma es la regla de cadena. La probabilidad de que los eventos A_1 *a* A_n, todos ocurran (la distribución de probabilidad conjunta) se denota como $p(A_1, A_2 \ldots A_n)$ y es igual a $p(A_n|, A_{n-1}, \ldots, A_1)$ x $p(A_n|, A_{n-2}, \ldots, A_1)$ x … x $p(A_2 \mid A_1)$ x $p(A_1)$. Por ejemplo, la propiedad de encontrar tres palabras en una secuencia es igual a la probabilidad de encontrar la tercera palabra dada la evidencia de las dos primeras palabras, multiplicada por la probabilidad de encontrar la segunda palabra dada la evidencia de la primera palabra, multiplicada por la probabilidad de la primera palabra.

Este tipo de análisis se ha empleado para combinar de manera óptima la información anterior a una palabra con la nueva evidencia provista por su ocurrencia, principalmente en reconocimiento de señales de voz [Rosenfeld, 94].

REESTIMACIÓN DE ESTADÍSTICAS

El valor de los modelos iterativos es que pueden emplearse aun cuando no hay una fórmula exacta para alcanzar una solución. En un procedimiento iterativo se hace una estimación inicial de la solución, y la estimación se prueba para ver si es aceptablemente próxima a la solución. Si no es así, la estimación se debe refinar. Las pruebas y las fases de refinamiento se repiten hasta que se alcanza una solución.

Ya que no se conoce una solución analítica para el problema de entrenamiento de los modelos ocultos de Markov, se pueden emplear técnicas iterativas, como el algoritmo de reestimación de Baum-Welch. La tarea es ajustar los parámetros del modelo para maximizar la probabilidad de una secuencia de símbolos observada [Sharman, 89].

Dado un modelo que produce una secuencia de símbolos observada, se quiere encontrar $\boldsymbol{x}_t(i,j)$, la probabilidad de que estando en un estado q_i en el tiempo t se haga una transición al estado q_j en el tiempo $t+1$.

$$x_t(i,j) = \Pr(i_t = q_i, i_{t+1} = q_j | O, l) = \frac{a_t(i) \times a_{ij} \times b_j(O_{t+1}) \times b_{t+1}(i)}{\Pr(O| l)}$$

donde:

$a_t(i)$ es la probabilidad de llegar al estado q_i en el tiempo t por cualquier trayectoria que salga del estado inicial, y produciendo el símbolo de salida O_t.

a_{ij} es la probabilidad de hacer la transición del estado q_i al estado q_j . Las probabilidades de la transición son parámetros originales del modelo.

$b_j(O_{t+1})$ es la probabilidad de producir el símbolo de salida en el siguiente paso O_{t+1}. Las probabilidades de la salida también son parámetros originales del modelo.

$b_{t+1}(i)$ es la probabilidad de dejar el estado q_j en el tiempo *t+1* por cualquier trayectoria, y eventualmente obtener el estado final.

La probabilidad de estar en el estado q_i en el tiempo *t* se llama $g_t(i)$ y se encuentra al sumar todos los valores de $\boldsymbol{x}_t(i,j)$ calculados para todos los valores de *i* desde 1 hasta N, el número total de estados en el modelo, como sigue:

$$g_t(i) = \sum_{i=1}^{N} x_t(i,j)$$

El número esperado de transiciones realizadas a partir del estado q_i se llama Γ_i, la cuál es la suma de todos los valores de $g_t(i)$ calculados en cada paso de tiempo desde *t=1* hasta *t=T* , donde *T* es el número total de pasos tomados por el modelo, como se muestra a continuación:

$$\Gamma_i(i) = \sum_{t=1}^{T-1} g_t(i)$$

El número esperado de transiciones realizadas del estado q_i al estado q_j se llama Ξ_{ij}, y es la suma de todos los valores de $\boldsymbol{x}_t(i,j)$ tomados en cada paso de tiempo desde *t=1* hasta *t=T* , como se muestra a continuación:

$$\Xi_{ij} = \sum_{t=1}^{T-1} x_t(i,j)$$

Con objeto de optimizar los parámetros del modelo para maximizar la probabilidad de la secuencia observada, se vuelven a estimar los valores de los tres parámetros que definen el modelo: las probabilidades iniciales del estado, las probabilidades de transición y las probabilidades de salida. Primero se reestima la probabilidad de cada uno de los estados iniciales. La probabilidad original del modelo estando en el estado *i* se llama p_i y la probabilidad reestimada se llama $\bar{p}_i$. Los valores $\bar{p}_i$ son iguales a los valores $g_1(i)$, los cuales son los valores cuando *t=1*. En segundo lugar, la nueva estimación de cada probabilidad del estado de transición, llamada $\bar{a}_{ij}$ se encuentra usando la relación

$$\bar{a}_{ij} = \frac{\Xi_{ij}}{\Gamma_i}$$

Ésta es la razón del número esperado de transiciones de un estado al siguiente, dividido por el número total de transiciones fuera de ese estado. Finalmente, la nueva estimación de cada probabilidad de salida, llamada $\bar{b}_i(k)$ es la razón de numero esperado de veces de estar en un estado y observar un símbolo, dividido por el número esperado de veces de estar en ese estado, dado por:

$$\bar{b}_i(k) = \frac{\Xi_{ij}(k)}{\Gamma_i}$$

Se tiene entonces un modelo nuevo $\bar{l}$, el cuál está definido por los parámetros de reestimación: $\bar{l} = (\bar{A}, \bar{B}, \bar{p})$. Estos valores pueden emplearse como los puntos de inicio de un nuevo procedimiento de reestimación, para obtener las estimaciones de parámetros que expliquen aún mejor la secuencia observada de símbolos.

Continuando este proceso iterativo, se llegará eventualmente a un punto donde los parámetros reestimados ya no son diferentes de los parámetros de entrada, es decir, los valores convergen. El punto de convergencia se llama un máximo local, que no impide la posibilidad de que el algoritmo pueda haber pasado por alto un mejor conjunto de parámetros llamado el máximo global. Si el máximo local es igual al máximo global, se encontró el conjunto de parámetros del modelo más próximo a explicar la secuencia observada de símbolos.

El algoritmo de Baum-Welch es un ejemplo de la clase de algoritmos denominados algoritmos de estimación-maximización (algoritmos EM), los cuales convergen en un máximo local. [Goodman, 98] presenta usos novedosos para las probabilidades interiores y exteriores, que se han empleado tradicionalmente para mejorar las probabilidades de las reglas de gramáticas CFG. Estas probabilidades se obtienen con algoritmos EM. La probabilidad interior es la probabilidad de que un no terminal consista exactamente de determinados terminales, es decir hacia dentro de su subárbol. Las probabilidades exteriores son las probabilidades de lo que está alrededor del no terminal, es decir, del contexto cercano. El producto de ambas probabilidades para los constituyentes de la oración da la probabilidad total de la oración con determinada estructura.

1.4 REDES SEMÁNTICAS

Existe una idea bastante extendida, tanto en la sicología como en la Inteligencia Artificial, de que en la mente humana los conceptos se encuentran relacionados entre sí formando una red. Bajo esta idea, cada concepto constituye un nodo de la red que se conecta con otros nodos mediante enlaces de distinta naturaleza. Los enlaces establecen el tipo de relación, entre ellos, algunos de los más empleados son: el enlace que indica pertenencia a una clase ("es un tipo de"), el de meronimia ("es una parte de"), el de sinonimia ("es igual que"), el de función ("tiene la función de"), el de contener ("contiene un"), etc. Este conjunto de nodos y enlaces se conoce como red semántica.

La red semántica es un conjunto de relaciones entre pares de palabras, o una combinación de palabras, refiriéndose a una cosa específica o idea. Si la palabra tiene diferentes sentidos, se incluyen en el diccionario en diferentes localidades y se marcan con diferentes números (por ejemplo *banco*$_1$, *banco*$_2$). Todos los sentidos de una palabra, aún los relacionados, tienen números diferentes y pueden conectarse explícitamente mediante relaciones. Así que una palabra puede representar muchos *conceptos* diferentes. De forma similar, un concepto puede representarse mediante varias palabras (*banco*$_1$, *taburete*, etc.), pero por conveniencia el concepto se marca con una sola palabra y no con el grupo de homónimos.

Como se observa en la Figura 14 una red semántica es un grafo. En ese grafo, hay cadenas de relaciones como las antes descritas. Una trayectoria se traza siguiendo las relaciones de una palabra a otra. De esta forma se puede medir que tan cercanas o lejanas en la red se encuentran pares de palabras. Dos consideraciones importantes deben tomarse en cuenta, primero que algunas relaciones solamente están presentes de forma implícita, por lo que se presenta el problema de generar todas esas relaciones aplicando reglas de inferencia. Segundo, algunas veces la relación entre pares de palabras no se puede establecer mediante alguna relación existente, por ejemplo un ser humano PUEDE_TENER un amigo que ES_CAPAZ de beber, entonces se deben emplear las dos relaciones.

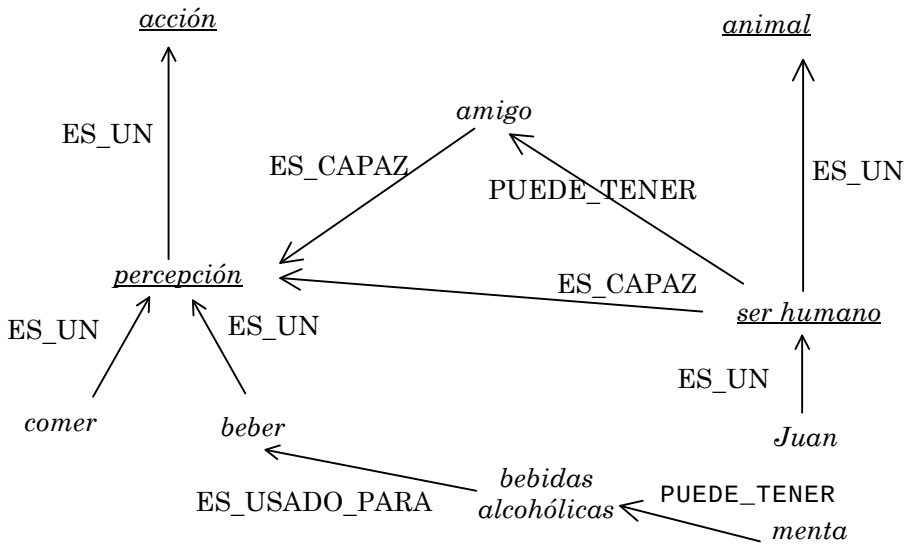


Figura 14 Red semántica para la frase
Juan bebe bebidas alcohólicas con sus amigos.

La dificultad que plantea este modelo simbólico es la delimitación de los diversos conceptos y de sus relaciones que intervienen en la red. Todavía se está muy lejos de poder establecer cuales son los conceptos básicos y de asignarles un contenido fijo. No hay por el momento un conjunto de conceptos o de primitivas semánticas universales. Por lo que cada grupo investigador tiene su conjunto de conceptos, aunque haya coincidencias entre ellos.

Aunque las redes semánticas también son una aproximación a las habilidades humanas y por lo tanto son modelos simplificados, pueden usarse de una forma acorde a sus limitaciones. Existen investigaciones en el área de lingüística computacional que han utilizado redes semánticas para resolver cierta clase de ambigüedad. Por ejemplo [Sanfilippo, 97] emplea WordNet [Miller, 90] para obtener automáticamente restricciones semánticas de ocurrencia concurrente para conjuntos de palabras relacionadas sintácticamente a partir de un corpus de textos. Propone crear clases de restricciones de ocurrencia concurrente, utilizando valores de entropía de los conceptos más informativos que incluyen en una categoría superior pares de palabras en la jerarquía.

Otro trabajo es el de [Rigau *et al*, 97], quienes proponen un método para desambiguar sentidos de palabras en un corpus grande sin marcas. Emplean diversas heurísticas, de entre ellas una es la *distancia conceptual* y la utilizan para determinar la cercanía entre significados de palabras. Esta distancia conceptual es la distancia

más corta que conecta los conceptos en la jerarquía. Emplearon WordNet como jerarquía y la distancia se mide entre la entrada de la definición del hipónimo y el *genus* de la definición del hiperónimo candidato.

En esta investigación también hacemos uso de la red semántica para reconocer cercanía de sentido entre grupos de constituyentes.

CAPÍTULO 2. COMPILACIÓN DEL DICCIONARIO DE VERBOS ESPAÑOLES CON SUS ESTRUCTURAS DE VALENCIAS

En este capítulo presentamos la caracterización de la estructura de valencias sintácticas con énfasis especial en las particularidades del español. Desarrollamos las características que se necesitan describir, presentamos estas características y sus valores para verbos y otras partes de la oración. Esta caracterización es finalmente la elaboración de las herramientas de la MTT para el análisis sintáctico del español, es decir el desarrollo de los patrones de manejo sintáctico para verbos principalmente, y también para sustantivos y adjetivos del español.

El problema de la caracterización es un problema de la lingüística general, pero aquí la consideramos más formalmente, desde el punto de vista computacional. Necesitamos términos y estructuras adecuadas para la descripción de los fenómenos lingüísticos. Las herramientas desarrolladas en la lingüística general se dirigen a los seres humanos, no a las computadoras, por lo que tenemos libertad para seleccionar medios semiformales. En la mayoría de los casos, tomaremos las estructuras usadas en la MTT, posteriormente mencionaremos otras maneras de descripción casi formal para enseguida presentar la comparación entre ellas.

2.1 DIVERSIDAD NUMÉRICA DE VALENCIAS

En los patrones de manejo sintáctico se describen todas las valencias de los verbos de acuerdo a su significado. En el español existen verbos con 0 valencias sintácticas y hasta 5, en general. Sin embargo, la mayoría de los verbos en español caen en el rango de 1 a 3 valencias. Algunas estadísticas acerca del número de valencias se presentaron en [Galicia *et al*, 98].

Comenzamos la explicación informal de los patrones de manejo, en el español, presentando algunos ejemplos de oraciones cuyos verbos tienen diferente número de valencias. Los números entre paréntesis indican las valencias semánticas y preceden a la correspondiente realización sintáctica.

Número de valencias	Ejemplos
0	Llueve.
1	(1) Juan duerme.
2	(1) Juan mira (2) las montañas.
3	(1) Juan acuerda (2) el proyecto (3) con su jefe.
4	(1) Juan compra (2) un vestido (3) en la tienda (4) en 500 pesos.
5	(1) Juan renta (2) un departamento (3) a María (4) por un año (5) en 500 pesos.

Aunque en la literatura que trata temas de constituyentes, se denomina *alternación* a la posibilidad para un verbo dado de aparecer en más de un tipo diferente de marco de subcategorización [Alsina, 93], que pueden relacionarse uno a

otro a través de un conjunto limitado de relaciones de mapeo, nosotros la denominamos *variación* ya que el término alternación se ha empleado principalmente como una noción morfológica [Alarcos, 84]. Ejemplos bien conocidos de estas variaciones son las siguientes:

- Existencia versus ausencia de objeto directo del verbo (su uso transitivo versus intransitivo). Por ejemplo *Juan comió un taco* y *Juan comió*.
- Agentiva versus instrumental. Por ejemplo: *Juan quebró el florero con el martillo* y *El martillo quebró el florero*. En el primer caso *martillo* es un instrumento, mientras que en el segundo, su función corresponde al agente que realiza la acción.
- Inversión. Por ejemplo: *Juan cargó la paja en la carreta* y *Juan cargó la carreta con paja*.

Mucho se ha escrito acerca de estas variaciones. [Atkins *et al*, 86] investigaron el rango de variaciones entre formas transitivas e intransitivas de verbos. [Levin, 93], [Levin & Rappoport, 91] exploran las relaciones entre el significado y las posibilidades de subcategorización para unos cuantos verbos, y sus agrupamientos relacionados. [Kilgarriff, 93] intenta generalizar el comportamiento de las variaciones en clases de verbos. Todos estos estudios consideran la clasificación de verbos, es decir, que se pueden agrupar diferentes verbos que presentan los mismos fenómenos sintácticos y que por lo tanto pueden analizarse en igual forma.

La clasificación de verbos se ha realizado desde diferentes perspectivas. En cuanto a estructura sintáctica se ha realizado considerando el tipo de complementos que son compartidos por diferentes verbos. Por ejemplo, en una forma simple, considerando el grupo de verbos transitivos cuyo objeto directo es un grupo nominal, o cuyo objeto directo es una frase preposicional, etc. [Kilgarriff, 92] presenta una clasificación de verbos más compleja, basada en conceptos semánticos y sintácticos en el nivel más alto de la jerarquía y con verbos de cierto tipo específico en los niveles inferiores.

Sin embargo, como se vio en la sección 1.2, en las consideraciones de la HPSG, algunos verbos con sentidos similares presentan diferentes marcos de subcategorización, además, como se verá más adelante, verbos con sentidos similares presentan diferentes números de valencias. Por lo que las clasificaciones, en cuanto a caracterización de valencia sintáctica y relación con la valencia semántica, no resulta ser la mejor forma de presentación.

En los patrones de manejo a diferencia de este tipo de clasificaciones, se describen individualmente las diferentes variaciones para cada verbo, lo que permite diferenciar las diferentes realizaciones sintácticas para cada verbo específico. En lugar de considerar diferentes grupos de variaciones como las mencionadas arriba, se analiza y describe individualmente cada verbo. Por ejemplo, establecer que el verbo

comer tiene una segunda valencia opcional que indica qué cosa se come.

Las mencionadas clasificaciones tienen la desventaja de no permitir esta diferencia ya que al agrupar formas de subcategorización pueden quedar en clasificaciones diferentes las variaciones para un verbo dado. Por ejemplo, el verbo *cargar*, tiene tanto la forma de subcategorización GN GP(*en*), *la paja en la carreta*, como GN GP(*con*), *la carreta con paja*, que denotan la inversión de los actuantes semánticos en el nivel sintáctico. Tampoco es posible apoyarse en esa clasificación para separar verbos homónimos cuando se detectan diferentes sentidos.

Al describir individualmente las diferentes valencias sintácticas de los verbos se describe el sentido implícito en las diversas construcciones. Para comparar con los roles semánticos presentamos el ejemplo clásico de *quebrar*, que se describe en los siguientes ejemplos de [Allen, 95]:

Juan quebró la ventana con el martillo.

El martillo quebró la ventana.

La ventana se quebró.

La tercera frase corresponde a la traducción del inglés de *The window broke*. Desde la perspectiva de roles temáticos, *Juan* es el actor (el papel del agente), *la ventana* es el objeto (el papel de tema) y *el martillo* es el instrumento (el papel de instrumento) usado en el acto de *quebrar*. La idea en los roles temáticos es generalizar los posibles participantes.

Desde el punto de vista de los actuantes semánticos, el sentido implicado requiere diferenciarse:

- En la oración *Juan quebró la ventana con un martillo*, una entidad animada utiliza un objeto para separar en dos o más partes otro objeto, con un fin determinado. Así que en la frase *Juan quebró la ventana*, está ausente el instrumento, que pudiera ser incluso su mismo cuerpo.
- En la oración *El martillo quebró la ventana*, un objeto separó en dos o más partes otro objeto; sin la participación de una entidad animada con un propósito específico.
- En la tercera oración *La ventana se quebró*, es una variante de se quebró la ventana, que indica la ausencia del objeto que separó en dos o más partes la ventana.

La descripción del verbo *quebrar* indicando sus valencias y posible separación en homónimos se presentará en Mapeo de valencias semánticas a sintácticas de la sección 2.8. De lo anterior, se desprende que el número de valencias sintácticas resulta de determinados rasgos semánticos, que deberán considerarse para caracterizar los verbos específicos. Este hecho se amplía en la sección 2.8

2.2 EJEMPLOS DE PATRONES DE MANEJO PARA VERBOS.

Aunque en los ejemplos de verbos que a continuación presentamos mencionamos la clasificación usual de transitivos e intransitivos también indicamos las diferencias que las valencias presentan.

Para todos los verbos del español en modo activo, la primera valencia o primer actuante es el sujeto gramatical de la oración, como se considera en la gramática clásica, y en forma simple se denomina sujeto[27]. Para muchos verbos, las características sintácticas del sujeto lo definen como un sustantivo animado. En la sección 2.4 daremos detalladamente las consideraciones de animidad en el español. Entonces, la lista de características sintácticas para el sujeto de esos verbos es (S, *an*), donde S indica sustantivo, y puesto que la marca de animidad sólo se da en sustantivos, esa lista podría simplificarse a (*an*).

La mayor parte de la información del sujeto en las oraciones en español, es común a casi todos los verbos, por lo que es mejor concentrarla en la gramática en lugar de situarla en los patrones de manejo sintáctico. Así que la descripción del sujeto, en los patrones de manejo sintáctico, normalmente se limitaría a la lista de índices léxico-semántico y semántico, como se verá en los ejemplos siguientes.

Verbos sin valencias

Los verbos españoles que no presentan ninguna valencia son los verbos que sólo se conjugan en tercera persona como: *llover, granizar, nevar*.

[27] Una excepción se presenta en los verbos que sólo se conjugan en la tercera persona del singular, los impersonales, como *llover*, *nevar*, etc. Estos verbos no tienen valencias en el nivel superficial, ni en el profundo.

llover

water falls to earth in drops

Solamente se considera su descripción.

Verbos con una valencia

Algunos verbos españoles intransitivos tienen únicamente la valencia que corresponde al sujeto, por ejemplo el verbo *cojear*:

cojear

person or animal X walks lamely

1 = X; quién cojea?

1.1 S % el hombre ~

% el gato ~

Por definición, los verbos intransitivos no pueden tener un complemento directo. Sin embargo, la ausencia del complemento directo es una peculiaridad puramente sintáctica. Los verbos intransitivos pueden tener otras valencias representadas mediante diversos complementos indirectos. Estas valencias, en el patrón de manejo, se numeran usualmente en el orden de importancia de ellas.

En los diccionarios comunes la información de las propiedades sintácticas de los verbos intransitivos de los lenguajes naturales, como el español, no considera estos posibles complementos. En el análisis sintáctico de textos por computadora es esencial esta definición para reducir la ambigüedad sintáctica. Por ejemplo, el verbo *perecer* tiene una segunda valencia realizada sintácticamente mediante un complemento indirecto que expresa la causa de la acción.

perecer

X ceases to live because of Y

1 = X; qué/ quién perece?

1.1 S % el hombre ~

2 = Y; ***de qué***?

2.1 de S % ~ de hambre

% ~ de frío

Estos verbos intransitivos tienen más de una valencia semántica y, en términos rigurosos, no pertenecen al grupo bajo consideración. Para marcar que algunos verbos intransitivos pueden presentar otras valencias, algunos diccionarios como LDOCE [Procter *et al*, 78] consideran la clasificación de intransitivos estrictos. Esta misma clasificación fue considerada en la descripción de subcategorización en las Gramáticas Categoriales (sección 1.2).

Verbos con dos valencias

Los verbos transitivos, por definición, tienen una segunda valencia semántica denominada en el nivel sintáctico como objeto directo o complemento directo. En muchas lenguas europeas, el complemento directo se une al verbo directamente, sin preposiciones. En español, existen dos posibilidades para esta conexión. Los sustantivos inanimados (*na*) generalmente se unen directamente al verbo, en cambio, los sustantivos animados (*an*) usualmente se unen al verbo mediante la preposición *a*.

Una característica de los verbos transitivos es que el complemento directo es obligatorio. Por ejemplo, la frase **Juan quiere* no es gramatical, requiere la indicación explícita de qué o a quién quiere *Juan*. Indicamos esta condición de obligatoriedad con el signo derecho de admiración (!) exactamente después de la fórmula de equivalencia entre valencias sintácticas y semánticas, por ejemplo: 2 = Y!; ***de qué***? Cuando la valencia no es obligatoria en el nivel sintáctico aparece únicamente el punto y coma.

A continuación presentamos un patrón de manejo sintáctico para el verbo transitivo *querer*:

querer$_1$

person X experiences positive feelings to person Y

1 = X; quién quiere?

1.1 (S, an) % el padre ~

2 = Y!; ***a quién***?

2.1 *a* (S, an) % ~ a su hijo

Verbos con tres valencias.

Los verbos considerados en la gramática clásica como doble transitivos tienen tres valencias. La tercera valencia se denomina objeto indirecto o complemento indirecto. En el español, los complementos indirectos siempre están unidos al verbo mediante preposiciones, por lo que frecuentemente se les denomina objetos preposicionales. Por ejemplo, el verbo *solicitar*:

solicitar

X asks something Y from Z

1 = X; quién solicita?

1.1 (S, an) % Juan / el gobierno ~

2 = Y!; qué solicita?

2.1 (S, na) % ~ una prórroga / un préstamo

2.2 *que* C % ~ que este libro se le dé

2.2 (V, inf) % ~ cancelar la autorización

3 = Z; de quién solicita?

3.1 *a* (S, an) % ~ a la secretaria

3.2 *con* (S, an) % ~ con el secretario

3.3 *de* (S, an) % ~ de usted

donde C indica una cláusula subordinada.

En el ejemplo *Juan solicita una prórroga al gobierno*, la primera valencia es el sujeto y la segunda el complemento directo. Para este verbo, se observa que además de que el complemento directo es obligatorio (no es posible decir **Juan solicita al gobierno*, sin indicar qué se solicita), existe un conjunto de preposiciones con las cuales la tercera valencia se une al verbo: *a*, *con*, y *de* (a diferencia de *perecer* donde la preposición *de* es la única que introduce la valencia).

La diferencia de significado de las frases *solicita un pase con el secretario* y *solicita un pase al secretario* no es de considerar, en la mayoría de los casos. Estas preposiciones no son sinónimas, su equivalencia está implicada en el verbo que las emplea. Otros verbos pueden usar un conjunto diferente de preposiciones para propósitos similares.

En la mayoría de los ejemplos previos las valencias se realizaron con sustantivos, pero las valencias pueden realizarse de formas diferentes. Por ejemplo, la segunda valencia del verbo *solicitar*, se puede realizar con *que* C.

Verbos con cuatro valencias

Otro ejemplo, el verbo *condenar* muestra el caso de una preposición que introduce un verbo en infinitivo:

condenar

person X condemns person Y to Z for action W

1 = X; quién condena?

1.1 (S, an) % el juez ~

2 = Y!; a quién condena?

2.1 *a* (S, na) % ~ al acusado

3 = Z; *a qué?*

3.1 *a* (S, na) % ~ a cadena perpetua

4 = W; por cuál motivo?

4.1 *por* (S, na) % ~ por asesinato

4.2 *por* (V, inf) % ~ por matar

En este ejemplo, la cuarta valencia presenta una forma diferente de realización de las anteriores, con un verbo en infinitivo, además de la forma más común, mediante un sustantivo.

En la gramática española [Seco, 72] se considera el hecho de que la forma del verbo refleja las distintas funciones que desempeña el núcleo de la oración; por ejemplo, cuando funciona como sustantivo, aparece en infinitivo. Para funcionar como adjetivo aparece en forma de participio, cuando funciona como adverbio, aparece como gerundio. Frecuencias de aparición de los distintos usos del infinitivo se encuentran en [Luna-Traill, 91], [Arjona-Iglesias, 91] y [Moreno, 85].

[Gili, 61] indica que mientras el francés desde el siglo XVI limitó mucho el número de infinitivos que pueden sustantivarse, el español ha conservado esta libertad, donde además se sustantiva la forma reflexiva. También indica que otras lenguas como el francés, el alemán y el inglés, limitan el número de preposiciones que pueden unirse al infinitivo, o bien restringen las construcciones verbales y sustantivas a que pueden aplicarse. Por lo que el empleo amplio de las preposiciones con los verbos en infinitivo es una peculiaridad más del español

El uso de preposiciones lleva aparejada ciertas dificultades. En algunos verbos, una frase preposicional describe tanto valencias del verbo como circunstancias. Por ejemplo, algunos verbos locativos [Rojas, 88] requieren complementos con la noción de espacio, cuya marca aparece tanto en la palabra introductora del complemento como en el complemento mismo. Por ejemplo, con el verbo *colocar*:

colocar$_1$

person X puts Y in place Z

1 = X; quién coloca?

1.1 (S, an) % el estudiante ~

2 = Y!; qué/ a quién coloca?

2.1 (S, na) % ~ los libros

3 = Z; dónde coloca?

3.1 *en/sobre* (S, na) % ~ en el estante

3.2 (Adv, loc) % ~ aquí

En la frase *coloca un libro en este momento en el espacio disponible*, la frase preposicional *en* NP describe tanto una valencia (*en el espacio libre*) como un complemento (*en este momento*) que es circunstancial de tiempo. Por lo que se requiere un descriptor como en el caso de animidad que distinga estos casos. Entonces

la tercera valencia se modifica a:

3 = Z; dónde coloca?

3.1 *en* (S, loc) % ~ en el estante

3.2 *sobre* (S, na) % ~ sobre la mesa

3.3 (Adv, loc) % ~ aquí

En donde *loc* indica sentido locativo del sustantivo.

Verbos con cinco valencias

Por último, presentamos un ejemplo, el verbo *rentar* que tiene cinco valencias:

rentar

person X uses the possession Y of the owner Z giving in return a quantityW by a period V

1 = X; quién renta?

1.1 (S, an) % María ~

2 = Y!; qué renta?

2.1 (S, na) % ~ un departamento

3 = Z; ***a quién?***

3.1 *a* (S, na) % ~ a la compañía Zeta

4 = W; en cuanto?

4.1 *en* (S, na) % ~ en dos mil pesos

5 = W; por qué período?

5.1 *por* S(tm) % ~ por mes

5.2 *a* S(tm) % ~ al mes

Donde *tm* significa tiempo y se refiere a un sustantivo que denota intervalo de tiempo.

2.3 EJEMPLOS DE PATRONES DE MANEJO PARA SUSTANTIVOS Y ADJETIVOS

Los adjetivos y los sustantivos difieren de los verbos, en las valencias que presentan, específicamente en la primera valencia. En los adjetivos, la primera valencia semántica es el correspondiente sustantivo que el adjetivo modifica. Desde el punto de vista semántico, los adjetivos son lexemas predicativos, que al menos tienen una valencia. El primer actuante es precisamente la palabra expresada mediante un sustantivo. En los verbos, la dirección de rección va del verbo al sujeto. En cambio en los adjetivos, la dirección de rección sintáctica es inversa, el arco de la relación sintáctica va del sustantivo al adjetivo. La razón es que el sustantivo es la palabra dominante y el adjetivo es la palabra dependiente.

El primer ejemplo que presentamos es un adjetivo homónimo: *blanco*. En español, existen al menos dos significados diferentes: *blanco*$_1$ con sentido referido al color, y *blanco*$_2$ con sentido referido a inocencia o pureza. Cada uno de estos homónimos tiene una sola valencia. Puesto que esta valencia no implica la correspondiente dependencia sintáctica, la fórmula 1 = X! tiene un carácter condicional y representa la llamada valencia pasiva. En este caso, la dependencia sintáctica (adjetivo de un sustantivo) es contraria a la dependencia semántica (sustantivo de adjetivo atributivo). Esta peculiaridad es inmanente a los adjetivos en muchos lenguajes, y desaparece al nivel semántico.

blanco$_1$

physical object X is of white color

1 = X!

1.1 (S, <Phys-obj>) % la pintura ~

blanco$_2$

narrative X is innocent or pure

1 = X!

1.1 (S, <Narr>) % chistes ~

Como usualmente la diferencia entre los homónimos se manifiesta en los descriptores semánticos de cada opción, <*Phys-obj*> denota un objeto del mundo y <*Narr*> denota el elemento del texto.

Los descriptores pueden emplearse para la desambiguación de los homónimos, cuando los dominios que ambos abarcan no intersectan. Por ejemplo, en la frase, *una pintura blanca, blanca* sólo puede referirse al color. En cambio, en la frase, *un libro blanco* hay duda del sentido asignado, el color del libro o su contenido.

Existen adjetivos con dos valencias. Por ejemplo, el adjetivo *lleno* en el que la segunda valencia expresa el objeto preposicional con el significado de establecer qué contiene en toda o casi toda su capacidad.

lleno

object X is full of object Y

1 = X!

1.1 (S) % el estadio ~

2 = Y; *de qué?*

2.1 *de* (S) % ~ de gente

El primer ejemplo que consideramos para los sustantivos, es un sustantivo que no deriva de forma verbal:

presidente

person X is the highest official of country or organization Y

1 = X; ***quién***?

1.1 (S, Propr) % ~ Adolfo López Mateos

2 = Y; de qué país u organización?

2.1 de (S, <Org>) % ~ de México, ~ del club

2.2 $\downarrow A_0$(<Org>) % ~ mexicano

El parámetro <*Propr*> denota una palabra o secuencia de palabras que expresan un nombre de persona; <*Org*> denota una subclase semántica de sustantivos que expresan un nombre oficial de una organización, la cuál puede ser de tipo social, político o cualquier otro tipo e incluye nombres de países.

La opción 2.2 es muy específica. El signo ↓ indica que la información de esta opción, en el nivel sintáctico, no se expresa mediante dependencia de valencia sino

por una de atribución, mientras que en un nivel más profundo la diferencia se elimina y la valencia semántica puede derivarse y representarse explícitamente. El término $A_0()$ es una función que deriva un adjetivo a partir del argumento, que es un sustantivo. Por ejemplo, A_0(*México*) = *mexicano*, A_0(*España*) = *español*, etc. En términos generales, esta opción presenta el ejemplo de una valencia semántica expresada por medio de otras dependencias sintácticas, ya que *México* y *España* son valencias semánticas.

Otro ejemplo con esta característica es el sustantivo *conclusión*, donde la opción 2.2 se expresa mediante adjetivo posesivo*:*

conclusión$_1$

It is the reasoned deduction or inference of person X on subject Y

1 = X; de qué persona?

1.1 *de* (S, <Person>) % ~ del profesor

1.2 ↓ (Adj, poss) % mi ~

2 = Y; sobre qué cosa?

2.1 *sobre* (S) % ~ sobre el proyecto

2.2 *de que* C % ~ de que el proyecto es...

2.3 *de* (S) % ~ de una serie de investigaciones

En este caso la marca de animidad se cambió por el descriptor semántico *persona* <Person>. En la mayoría de las ocasiones ambas etiquetas significan lo mismo, seres humanos. Pero la animidad se aplica también a entidades personificadas como animales, grupos de personas, países, etc. y realmente el sentido es más estrecho en este caso donde no es posible imaginar su uso en el ámbito extendido de personificación. Aunque por otro lado, en el mundo contemporáneo, una conclusión también puede realizarla un autómata que razone. El descriptor semántico corresponde al ámbito bastante estrecho de este caso particular. Puede ocurrir que el descriptor de este tipo no “funcione” correctamente en casos de métafora.

El último ejemplo que presentamos es el sustantivo *querella*, que generalmente no va acompañado de adjetivos.

querella$_1$

complaint of person X against person Y on subject Z

X = 1; de quién?

1.1 *de* (S, an) % del vecino ~

1.2 ↓ (Adj, poss) % mi ~

Y = 2; contra quién?

2.1	*contra* (S, an)	% ~ contra quién resulte responsable
2.2	*en contra de* (S, an)	% ~ en contra de Juan

Z = 3; por qué?

3.1	*por* (S, na)	% ~ por robo
3.2	*por* (V, inf)	% ~ por defraudar
3.3	*de* (S, na)	% ~ de robo

La segunda valencia sintáctica presenta una de las peculiaridades de muchos lenguajes modernos: se trata de una preposición compuesta. Además de las preposiciones comunes simples que registran los diccionarios como tales, existen numerosas *locuciones preposicionales* o *locuciones prepositivas* en las cuales figuran ordinariamente un sustantivo o un adjetivo: *alrededor de*, *encima de, dentro de, junto a, frente a, enfrente de*, etc. y otras muchas que ocasionalmente pueden crearse para precisar la relación, a veces poco definida, de las preposiciones solas. De esta manera, y con la combinación de dos o más preposiciones, el español compensa el número relativamente escaso de preposiciones simples.

Algunas de estas locuciones prepositivas son casi del todo equivalentes a preposiciones simples, y en ocasiones más usadas que éstas: *delante de* (= ante), *encima de* (= sobre), *debajo de* (= bajo), *detrás de* (= tras). También el adverbio se suma a la función enlazadora aportada por la preposición, y la unión de las dos palabras se convierte en una nueva preposición: *antes de*, *encima de*.

Se forman nuevos conjuntos uniendo las preposiciones a otras preposiciones, dando lugar a complejos característicos del español, en los que la aglomeración de preposiciones expresa una gran variedad de relaciones. Por ejemplo: ***de entre** ellos*, ***de con** sus amigos*, ***desde por** abajo*, ***hasta con** sus compañeros*, ***para entre** nosotros*, ***por de** pronto*. A veces llegan a reunirse hasta tres preposiciones, por ejemplo: ***hasta de con** sus compañeros fueron a buscarla*; ***desde por entre** los árboles nos espiaban*. Estos grupos se consideran como una sola preposición introductora de realizaciones sintácticas en los patrones de manejo.

2.4 DEPENDENCIA DEL OBJETO DIRECTO EN LA ANIMIDAD, COMO UNA PECULIARIDAD DEL ESPAÑOL

En la mayoría de los lenguajes el objeto directo está conectado con el verbo directamente, sin preposiciones, es por esto precisamente que este objeto se denomina directo. Por el contrario, en español, las entidades animadas están conectadas a su verbo rector con la preposición *a* (*veo a mi vecina*) y las no animadas directamente (*veo una casa*). La animidad en español se considera como una personificación. Por ejemplo, *gobierno* en español es un sustantivo animado y al dirigirse a él se utiliza la preposición *a* (*condenó al gobierno*). Además de personas, la animidad abarca grupos de animales, personas, organizaciones, partidos políticos, países, etc.

En otros lenguajes, por ejemplo el ruso, también existe una categoría de animidad similar que determina la terminación del caso morfológico del complemento directo, pero los grupos de personas, los países, las ciudades no se personifican en sentido gramatical[28].

Entonces la regla léxica del español es que el complemento directo está unido al verbo rector a través de la preposición *a* para entidades animadas y directamente en los otros casos. Su empleo es específico del español y lo distingue de otros lenguajes europeos. Pero la animidad en español realmente es aún más complicada, ya que en algunos contextos de indefinición o de conteo el complemento directo animado puede eliminar la necesidad de la preposición, por ejemplo: *Vio un niño que corría; Necesita tres ayudantes*. El contexto influye en algunos casos para incluir la conexión de objeto directo o eliminarla. Por ejemplo, un animal es animado en un ámbito de relación cercana al hablante: *veo a mi perro adorado* y es no animado en un ámbito

[28] En ruso se consideran como animados a la par de los seres humanos y animales, las muñecas, los insectos, etc.

de relación lejana: *veo el perro que corre por la pradera*.

Por el orden de palabras no estricto del español, se presentan combinaciones donde existe ambigüedad para detectar el objeto directo. Por ejemplo la frase *la realidad supera la ficción* podría presentarse en la forma verbo, sujeto, objeto directo (**supera la realidad la ficción*). Para comprender la frase correctamente, es decir, para identificar los argumentos, se emplea la preposición *a* antes de objetos inanimados: *supera la realidad a la ficción.*

Entonces, la categoría de animidad tiene en el español, dos valores opuestos: *an* (animado) y *na* (no animado). En el siguiente patrón de manejo mostramos las distintas formas en que se presenta la marca de animidad.

atrapar$_1$

X using force catches Y

1 = X; quién atrapa?

1.1 (S, an) % el policía / el gato ~

2 = Y!; qué / a quién?

2.1 (S, na) % ~ la maleta

2.2 *a* (S, an) % ~ a un ladrón

En la primera valencia se expresa el uso de sustantivos animados, en la segunda valencia se expresa el objeto directo con sustantivos no animados (S, na) y con la marca de animidad (S, an). Esta última responde precisamente a la pregunta ¿a quién? que se aplica tanto a una persona o a un animal, como a un grupo, a un partido, etc. Por ejemplo: *¿A quienes atrapó la policía*? y la respuesta: *al equipo de fútbol de la secundaria 28*.

Así que la animidad es una característica evidentemente sintáctica pero con alusión semántica que se considera para la realización de las valencias. Su principal importancia es la característica gramatical del español al conectar el objeto directo animado con preposición, aunque como hemos visto su uso puede ser útil para otros casos.

2.5 OTRA DEFINICIÓN DE LA NOCIÓN DE ANIMIDAD Y SU USO

Si definimos la noción de animidad en un sentido puramente semántico, como una característica de los seres vivos, entonces comprenderíamos de inmediato que hay una gran diferencia entre esa noción semántica y la animidad gramatical. La animidad semántica, entonces, sólo debería tomarse como característica de valencias de verbos orientados a "lo humano" como leer, cuyo sujeto puede ser solamente un ser humano o un autómata inventado en este último siglo con ese fin específico. En otros casos deberá ser únicamente asociada a los seres humanos ya que es difícil, al menos en este tiempo, asociar un autómata con verbos como: procrear, morir o imaginar.

Por supuesto que no consideramos ni las metáforas ni la poesía como ¡*Canta, lluvia, en la costa aún sin mar*![29] ya que la comprensión de ellas será posiblemente comprendida por las computadoras, en siglos futuros.

Ya que la noción de animidad en el sentido semántico de criaturas vivientes no incluye las entidades personificadas, se requiere una mayor investigación para definir exactamente si en todas las aplicaciones la característica de animidad es la misma característica de animidad del español o una muy similar a la característica semántica de seres vivientes. Por lo que de aquí en adelante emplearemos únicamente la característica de animidad gramatical para otras valencias.

Es conocido que la preposición *a* también tiene otros usos, no relacionados con la conexión del objeto directo. Pero aún para el objeto directo su empleo puede servir para diferenciar el significado de algunos verbos, por ejemplo, *querer algo* (tener el deseo de obtener algo) y *querer a alguien* (amar o estimar a alguien). El primero ya se presentó en la sección 2.2 como $querer_1$ y el segundo corresponde a:

[29] Cesar Vallejo "Trilce", fragmento del LXXVII

querer$_2$

person X desires thing Y

1 = X; quién quiere?

1.1 (S, *an*) % el niño ~

2 = Y!; ***qué***?

2.1 (S, *na*) % ~ un triciclo

Otra utilidad del empleo de la animidad, está relacionada con la primera valencia de algunos verbos, es decir, con el sujeto, y con su detección para reconocer valencias sintácticas. Por ejemplo, el verbo *acusar*, tiene dos homónimos: *acusar*$_1$ (denunciar a alguien como culpable de algo) y *acusar*$_2$ (revelar algo, ponerlo de manifiesto) conforme a [DEUM, 96]. Presentamos a continuación sus patrones de manejo:

acusar$_1$

person X accuses person Y of action Z

1 = X; quién acusa?

1.1 (N, *an*) % el ministro ~

% la madre ~

2 = Y!; a quién acusa?

2.1 *a* (N, *an*) % ~ a los políticos

% ~ a los niños

3 = Z; ***de qué?***

3.1 *de* (N, na) % ~ de robo

3.2 *de* (V, inf) % ~ de defraudar

acusar$_2$

X reveal Y

1 = X; quién/ qué acusa?

1.1 S % el ministro ~, la puerta ~

2 = Y!; ***qué acusa?***

2.1 (S, na) % ~ cansancio

% ~ el paso de los años

En algunas construcciones de *acusar*$_1$ como las siguientes, el sujeto aparece pospuesto al verbo en dos formas: como nombre propio y como nombre común:

- *Al director le acusaba Apel de desembocar en una ilusión idealista por ocuparse de <...>.*
- *En el presunto fraude aparece como principal sospechoso José Joaquín Portuondo, a quien acusaron varios testigos.*

En la primera frase el reconocimiento de nombre propio permite identificar el sujeto. En la segunda frase se requiere reconocer la entidad animada marcada en la realización del sujeto de *acusar*$_1$, es decir, reconocer que *varios testigos* es el sujeto. De esta forma no habrá confusión con el objeto de *acusar*$_2$, por ejemplo en oraciones donde las realizaciones de las valencias no son muy diferenciadas: *Que acusaba un alto rendimiento, le acusaba un alto magistrado*. Esta confusión puede resultar en una asignación de estructura incorrecta o en detección de valencia de otro sentido.

2.6 REPETICIÓN LIMITADA DE LOS OBJETOS COMO OTRA PECULIARIDAD DEL ESPAÑOL.

Generalmente las entidades referidas por las diversas valencias sintácticas son diferentes. Ésta es una situación normal en los lenguajes naturales: cada valencia semántica se puede representar en el nivel sintáctico por solamente un actuante.

Pero existen lenguajes en los cuales se permite la repetición restringida de actuantes. El español se cuenta entre esos lenguajes. En las frases siguientes, en cada oración, los dos segmentos disjuntos marcados con negritas se refieren al mismo objeto:

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

Mientras en la primera frase se repite el objeto indirecto, en las dos últimas frases se repite el objeto directo.

Algunas veces la repetición es obligatoria. El orden de palabras y los verbos específicos imponen ciertas construcciones. Por ejemplo, la anteposición de los argumentos dativos y acusativos presenta una complicación.

Las siguientes frases con objeto directo no permiten la duplicación:

Arturo escribió la carta.

Escribió Arturo la carta.

En cambio la anteposición del objeto directo (**La carta escribió Arturo.*) requiere o una marca de puntuación o la duplicación para ser correcta:

La carta, Arturo la escribió.

La carta la escribió Arturo

Para el objeto directo animado, que se introduce con la preposición *a*, tenemos los siguientes ejemplos:

El frío mató a la mosca.

Mató el frío a la mosca.

Mató a la mosca el frío.

Pero la anteposición del objeto directo animado (**A la mosca mató el frío.*) requiere la duplicación para ser correcta:

A la mosca la mató el frío.

En el ejemplo para objeto indirecto que a continuación mostramos, se permite la duplicación:

Arturo escribió una carta a Víctor.

Arturo le escribió una carta a Víctor.

También la anteposición del objeto indirecto (**A Víctor escribió una carta Arturo.*) requiere la duplicación para ser correcta:

A Víctor le escribió una carta Arturo.

[Zubizarreta, 94] afirma que la existencia de objetos doblados por clíticos es una diferencia más del español, ya que no existe en el italiano escrito, ni en el francés ni en algunos otros lenguajes europeos.

Notamos, por el contrario, que los siguientes ejemplos no están relacionados con la repetición de objetos, corresponden a otra condición. Mientras en la primera frase se omitió la tercera valencia del verbo *ordenar* (a quién se ordenó algo), en la segunda frase se representó con *le*.

El juez ordenó tomar declaración al acusado.

El juez le ordenó tomar declaración al acusado.

El objeto indirecto *a nadie* también puede repetirse con el clítico dativo pero dentro de su propia cláusula, no puede moverse a cláusulas superiores [Zubizarreta, 94]. Por ejemplo:

A nadie le dijo Juan de su boda.

**A nadie piensa María que le dijo Juan de su boda.*

Los pronombres personales también se pueden repetir:

A todos les dijo Juan de su boda.

En todos los casos la repetición de actuantes, se restringe mediante las siguientes reglas:

- Uno de los actuantes repetidos es un pronombre personal en caso indirecto (acusativo o dativo). Para las formas personales este pronombre

usualmente permanece justo antes del verbo rector, en los infinitivos se pega al verbo.

- Otra repetición del mismo actuante se da con un sustantivo, pronombre o algún otro medio. Pero en el caso de un pronombre personal, se pone en nominativo, por ejemplo: *A ellas las encontrarás en la tienda*.

En el nivel semántico de representación, todos los casos de esas repeticiones deben juntarse, es decir, se debe dejar un solo representativo de cada actuante semántico.

2.7 EL COMPLEMENTO BENEFICIARIO EN EL ESPAÑOL Y SU DUPLICACIÓN

En la caracterización semántica de complementos de muchos verbos en varios lenguajes, se considera la noción importante de la persona que recibe el interés, la beneficiaria, y la persona destinataria. El interés incluye los sentidos de daño y provecho. Para nuestro estudio son importantes los siguientes aspectos:

- Las personas beneficiaria y destinataria están representadas por el mismo objeto indirecto, como en *enseñar algo a alguien*, donde *a alguien* es tanto el beneficiario como el destino. En este caso hay que hacer notar que el complemento beneficiario (o benefactivo) corresponde a una valencia semántica del verbo.
- El beneficiario de la acción no corresponde a ninguna valencia. Es como una circunstancia. Usualmente se introduce mediante la preposición *para*. Por ejemplo: *comprar un libro para alguien*, donde el beneficiario se representa con un complemento circunstancial del verbo.

Claro que la preposición *para* puede introducir también otro tipo de complementos. Por ejemplo, en la frase *todo esto es para que te acostumbres* (con el sentido de meta), *habla muy bien para la edad que tiene* (con el sentido de comparación).

Algunos verbos llevan entonces el complemento beneficiario con cualquiera de las preposiciones: *a* o *para,* por ejemplo:

Víctor concedió una entrevista a la revista C&S.

Víctor concedió una entrevista para la revista C&S.

En estas frases puede haber una ligera discrepancia en el sentido, si la acción se realizó directamente o indirectamente. Sin embargo el beneficiario y el destino coinciden. Otras frases pueden mostrar ambigüedad de sentido con esta alternancia de preposiciones:

Emma escribe una carta a Emilio.

Emma escribe una carta para Emilio.

En esta sección nos concentramos en el complemento beneficiario del verbo predicativo y no en otros beneficiarios expresados en la oración, por ejemplo:

Paco dio un libro a su ayudante para Emilio.

Donde *para Emilio* es complemento beneficiario de *un libro* y el complemento que recibe el interés del verbo *dar* es el complemento *a su ayudante*, es decir, *Paco dio a su ayudante un libro para Emilio*.

Introducimos una clasificación de verbos de acuerdo a características transitiva, dativa y beneficiaria.

En esta sección, los verbos del tipo 0 y 1, sin beneficiario potencial, no son relevantes. No existen verbos con características dativa y beneficiaria distintas, por lo que no consideramos los tipos 6 y 7. A continuación presentamos ejemplos de cada tipo:

Del tipo 2 son: *corresponder, competer, convenir, doler, parecerse, gustar, faltar*, etc. Ejemplos de frases: *La creación de leyes corresponde al gobierno, Este asunto compete a todos*. Con repetición dativa: *La creación de leyes* ***le*** *corresponde al gobierno, Este asunto* ***les*** *compete a todos*.

Tipo	Transitivo	Dativo	Beneficiaria	Nota
0	—	—	—	No relevante
1	+	—	—	No relevante
2	—	+	—	
3	+	+	—	
4	—	—	+	
5	+	—	+	
6	—	+	+	No existen
7	+	+	+	No existen

Del tipo 3 son los más comunes: *dar, enseñar, asignar, preguntar, costar, interesar, servir* (algo a alguien), etc. Ejemplos de frases: *Alberto enseña anatomía a sus alumnos, Alberto preguntó todos los pormenores al abogado*. Con repetición

dativa: *Alberto **les** enseña anatomía a sus alumnos, Alberto **le** preguntó todos los pormenores al abogado*.

Del tipo 4 son: *laborar, cabildear*, etc. Ejemplos de frases: *Rodrigo labora para la compañía X*. *Los políticos cabildean para sus amos*. Con repetición beneficiaria: no se encontraron.

Del tipo 5 son: *comprar, componer, comprobar*, etc. Ejemplos de frases: *Emilio compone los audífonos para su hermano*. *Arturo comprueba los resultados para su jefe*. Con repetición beneficiaria: *Emilio **le** compone los audífonos **a** su hermano*. *Arturo **le** comprueba los resultados **a** su jefe*.

Existe una larga discusión acerca de la naturaleza argumental del complemento beneficiario. Entre los autores, [Branchadell, 92] para el español y [Jackendoff, 90] para el inglés, consideran que los beneficiarios no son valencias pero se comportan como ellos, por esto los podemos denominar *casi* valencia.

En el español, el punto de vista de consideración como casi valencia está bien fundada ya que a veces se duplica. Pero esta duplicación está sujeta a ciertas restricciones de realización léxica. Se realiza mediante un pronombre personal en la forma clítica o mediante un grupo preposicional. Por ejemplo:

*Emma **le** ha reservado unos lugares **a su familia**.*

En esta frase tanto *le* como *a su familia* representan la cuasi valencia beneficiaria, de la misma forma que *me* y *a mí* en el siguiente ejemplo:

*Emma **me** ha reservado unos lugares **a mí**.*

Sin embargo, la frase:

Emma ha reservado unos lugares para su familia.

no permite esa clase de repetición. Otros ejemplos presentados por [Demonte, 94] explican cómo la cuasi valencia beneficiaria no es posible en ciertas estructuras, por ejemplo:

Le coloqué cortinas al salón.

**Coloqué cortinas para el salón.*

**Le coloqué cortinas para el salón.*

Y tampoco es posible en ciertos contextos, por ejemplo:

Le puse el papel a la pared.

**Le puse un clavo a la pared.*

La autora considera que es por razones del mundo, la pared forma una unidad con el papel y el clavo no.

En otros casos puede verse claramente que existe una distinción entre el complemento indirecto y el destinatario, comparando la diferencia de significado que presentan las dos frases siguientes:

***Le** di un mensaje **para ti**.*

***Te** di un mensaje **para él**.*

En los ejemplos, *le* y *te* son los complementos indirectos, en cambio *para ti* y *para él* son los complementos de destinatario. Si todos los complementos subrayados fueran indirectos, no habría diferencia de contenido entre las dos oraciones.

Esta casi valencia con duplicación potencial se requiere describir explícitamente en los patrones de manejo de los verbos de los tipos 4 y 5. Sin su descripción estaría incompleta la relación con las valencias semánticas. Para los casos de duplicación, su descripción es necesaria para juntarlas en el nivel semántico.

2.8 OTRAS COMPLEJIDADES DE LA REPRESENTACIÓN DE VALENCIAS

Todas las valencias semánticas son necesarias y suficientes, y el papel que cada una de ellas desempeña está implicado directamente por la definición lexicográfica del lexema correspondiente. La situación con las valencias sintácticas es mucho más complicada y no siempre existe una correspondencia de uno a uno entre valencias sintácticas y semánticas. Se presentan los siguientes casos:

1. Estado incompleto en el nivel sintáctico. Frecuentemente las valencias semánticas no se reflejan en un texto dado, aunque sea totalmente gramatical.
2. Correspondencia desigual entre valencias sintácticas y semánticas:
   - Dos valencias semánticas pueden implementarse en una valencia sintáctica común.
   - Una valencia semántica puede implementarse en dos valencias sintácticas que concurren en un texto.
3. El mapeo de valencias semánticas a sintácticas puede sujetarse a corrimientos y a permutaciones, específicos.

Estado incompleto en el nivel sintáctico

El estado incompleto en el nivel sintáctico se restablece fácilmente (si se necesita) en el cerebro humano, y usualmente ni se menciona. Como ya vimos, existen las valencias sintácticas obligatorias, pero pueden no cubrir todas las valencias. Por ejemplo, si un hablante quiere informar a alguien, de una compra, pero para él son irrelevantes el vendedor y el precio, en la comunicación que se lleva a cabo, el hablante incluirá solamente en su expresión palabras correspondientes al comprador y a la compra. Las valencias semánticas del vendedor y el precio no las

realiza sintácticamente, por ejemplo: *Javier compró un libro*.

Es muy frecuente que no aparezcan en una oración todas las realizaciones sintácticas de las valencias del verbo, entonces se pueden ordenar las valencias semánticas de las palabras por la importancia de sus valencias sintácticas. El orden comienza entonces con las valencias absolutamente obligatorias.

La ocurrencia de valencias para algunas palabras no puede describirse mediante leyes simples, además una tendencia válida para un lenguaje no es verdadera para otro. Por ejemplo, en inglés se puede decir *It depends,* y en español *Depende,* cuando una situación depende de algunas circunstancias que se describirán posteriormente. En cambio, en ruso la oración equivalente a esta última es imposible. Esto significa que la segunda valencia sintáctica es optativa para este verbo en inglés y en español, pero no en ruso. Por lo que la obligatoriedad de valencias sintácticas es una cuestión particular de un lexema específico en un lenguaje específico.

Correspondencia desigual entre valencias sintácticas y semánticas

La correspondencia desigual entre valencias sintácticas y semánticas, la dividimos en dos casos:

1. Dos valencias semánticas pueden implementarse en una valencia sintáctica común, por lo que el número explícito de valencias sintácticas puede ser menor en el nivel semántico. Por ejemplo, el verbo *discutir*:
    - Los profesores discutieron con los funcionarios del Ministerio.
    - Los profesores discutieron.

En estas dos frases se observa que una valencia ausente se realiza implícitamente en la segunda oración. En la primera oración, las dos valencias sintácticas que corresponden a las valencias semánticas separadas: *¿quién está replicando?* y *¿con quién?* Mientras que en la segunda frase, la valencia *¿con quién?* está explícitamente ausente. Pero la situación difiere bastante del estado incompleto considerado anteriormente. De hecho, la valencia aparentemente ausente está realizada implícitamente también en la segunda valencia pero como una sobrecarga del sustantivo expresado en la primera valencia. Los profesores replican *uno con otro*, es decir, entre ellos, así que ambas valencias están presentes, una de ellas en esta forma tan particular.

En el caso del verbo *discutir*, el sentido mismo de por sí implica más de un participante; sin embargo también puede estar presente este fenómeno en el verbo *hablar*:

Los profesores hablan con los funcionarios del Ministerio.

- *Los profesores hablan.*

Aunque en la segunda frase, considerar que los profesores hablan entre ellos, sólo está relacionado al sentido de conversar y no al de pronunciar sonidos.

En español tanto como en otros lenguajes, existen varios verbos relacionados con información y actividades de comportamiento, recíprocas. La ausencia de la correspondiente valencia sintáctica se trata como la equivalencia entre el actuante faltante y el primero. El patrón de manejo para esos verbos se complica ya que se requiere adicionar una fórmula condicional especial.

IF (3 = ∅) & (1 = (N, *Mult*)) THEN (3 ⇐ 1)

Esta condición tiene la siguiente interpretación: si el tercer actuante se omite en el texto, y el primer actuante es un sustantivo con la propiedad *Mult*, es decir, es plural o expresa una multitud (grupo), entonces el tercer actuante se establece igual al primero. Esta condición involucra considerar a la tercera valencia como obligatoria.

2. Una valencia semántica puede implementarse en dos valencias sintácticas que concurren en un texto.

Un ejemplo de este caso corresponde al fenómeno presentado en la sección 2.6, donde se repiten valencias sintácticas que deben concentrarse en un sólo representativo de la correspondiente valencia semántica.

Mapeo de valencias semánticas a sintácticas

El tercer y último caso corresponde al mapeo de valencias semánticas a sintácticas, que puede sujetarse a corrimientos y permutaciones, específicos. Cada corrimiento o permutación implica una nueva versión de los PM para el lexema predicativo. Presentamos primero un ejemplo de corrimientos y después uno para permutaciones.

CORRIMIENTOS

El ejemplo bien conocido del verbo *quebrar* (sección 2.1), considerado como ejemplo clásico de roles temáticos, lo consideramos ahora para explicar los corrimientos. La concepción común de *quebrar* puede explicarse mediante la siguiente definición: *Person X split into two or more pieces a thing Y with instrument Z*, a esta descripción corresponde la frase *Juan quebró la ventana con el martillo*. El patrón de manejo considerará entonces las valencias ¿quién quiebra?, ¿qué quiebra? y ¿con qué quiebra?

La frase *El martillo quebró la ventana*, contiene entonces un corrimiento metafórico de valencias, la valencia Z se vuelve la primera valencia sintáctica y desaparece la valencia X. la valencia Y sigue siendo la segunda valencia sintáctica. Así que se puede introducir este último patrón de manejo como

quebrar$_1$

object Z divides into two or more pieces an object Y.

1 = Z; ***qué***?

1.1 N

2 = Y! ***qué***?

2.1 N

Y en base a este patrón de manejo definir el homónimo:

quebrar$_2$

Person X causes Z ***quebrar***$_1$ *Y.*

1 = X; ***quién***?

1.1 N

3 = Z; ***con qué***?

3.1 *con* N

De esta forma ***quebrar***$_2$ está en acuerdo total con la descripción de ***quebrar***$_1$. La mejor aproximación sería la interdependencia de homónimos. La aproximación con dos homónimos parece preferible para los lingüistas.

PERMUTACIONES

Las permutaciones consideradas de valencias ocurren porque el sentido del verbo es recíproco o porque la realización sintáctica las permite.

El primer caso se presenta con algunos verbos cuyo sentido permite que sean posibles las permutaciones. Por ejemplo, con los verbos *contrastar* y *rimar*. Sus valencias pueden intercambiarse sin mayor problema, y sus realizaciones sintácticas son iguales, no hay diferencia.

La sequedad del campo contrasta con la vegetación del jardín.

La vegetación del jardín contrasta con la sequedad del campo.

“Sala” rima con “pala”.

“Pala” rima con “sala”.

En los ejemplos anteriores no se requiere diferenciar valencias porque son iguales. La comparación de los siguientes ejemplos para el verbo *abundar*, indica que se requiere crear una definición semántica para cada realización.

El río abunda en peces

Los peces abundan en el río.

En ambos casos se presenta algo como: *espacio X contiene una gran cantidad de objetos Y.* Pero comparando con las frases anteriores, las dos posibles alternativas

de los patrones de manejo tienen evidentemente valencias permutadas, una contra otra.

La primera alternativa, de un patrón de manejo, correspondiendo al primer ejemplo para *abundar* es:

1 = X; cosa que contiene?

1.1. (N, Loc) % el río ~ / el bosque ~

2 = Y! de qué / de quién?

2.1 de (N, Mult) % ~ de frutas / de gente / de peces

2.2 en (N, Mult) % ~ en frutas / en gente / en peces

La etiqueta *Loc* se asigna a un sustantivo con propiedad de espacio. La etiqueta *Mult* significa el sustantivo en plural (como *flores, bosques, peces, ideas*) o que tiene el significado de colección (como *gente*).

La segunda alternativa de un PM, correspondiendo al segundo ejemplo para *abundar* es:

1 = X; qué está contenido?

1.1. (N, Mult) % los peces ~

2 = Y! ***dónde***?

en (N, Loc) % ~ en el río

La situación puede describirse entonces a través de dos verbos homónimos. De estos ejemplos se infiere que la aplicación lingüística requiere la descripción completa de todas las posibles opciones.

2.9 EJEMPLOS DE COMPLICACIONES DE PATRONES DE MANEJO PARA VERBOS DEL ESPAÑOL

Los actuantes deben tomarse en una cantidad necesaria y suficiente. Pero las valencias no son entidades predeterminadas y aún en palabras sinónimas pueden diferir en su significado y número. Por lo que la descripción de algunos patrones se vuelve complicada al tratar de establecer cuáles son las valencias necesarias y suficientes.

La descripción de cuantos actuantes están implicados en el lexema está relacionado a su sentido pero esto no quiere decir que los grupos de verbos con significado semejante tengan el mismo número de actuantes. Por ejemplo los verbos *decir* y *contar* que tienen un sentido similar, tienen diferente número de valencias, entre los distintos sentidos presentamos los siguientes:

decir$_1$ Alguien X dice algo Y a alguien Z sobre algo o alguien W

Por ejemplo: *Juan dijo cosas horribles a María del profesor*.

contar$_1$ Alguien X cuenta algo Y a alguien Z

Por ejemplo: *Juan contó una historia a sus alumnos*.

Las valencias semánticas no están predeterminadas y difieren aún para lexemas sinónimos. Por lo que el marcado de valencias semánticas requiere un trabajo manual. Los verbos *decir* y *contar*, que tienen un sentido similar pero presentan diferente número de valencias. Las oraciones con *decir* pueden contener la frase referente al tema (sobre algo o alguien), mientras que las oraciones para *contar* rara

vez la contienen.

Por ejemplo, la definición para el verbo español ***acusar***$_1$ podría empezar con *X acusa a Y de Z*, y a partir de ella tratar de explicar la situación con palabras adecuadas que reflejen todas las situaciones. Los tres actuantes involucrados son las valencias semánticas necesarias y suficientes.

Pueden implicarse otros vínculos, como razonamiento, tiempo y lugar. Por ejemplo: *Juan acusa a María en base a pruebas irrefutables*, *Juan acusó a María en el mes de mayo* o *Juan acusa a María en los tribunales federales*. Sin embargo, ninguna de estas circunstancias es obligatoria para distinguir entre **acusar**$_1$, y otros verbos con sentido similar, como culpar, condenar, imputar, etc. por lo que se comprueba que las valencias descritas son las necesarias y suficientes.

La definición del número de valencias puede complicarse. Existen verbos para los cuales es complicado determinar el número exacto de valencias. Por ejemplo, el verbo *enviar* se emplea frecuentemente con objetos de información (cartas, mensajes, archivos, etc.), los cuales se envían a través de una oficina de correos, un transmisor, un canal de comunicación, etc. Así que, con este sentido, es válido asignar al verbo *enviar* cuatro valencias: ¿quién o qué envía?, ¿qué se envía?, ¿adónde se envía?, ¿vía que canal? Entonces, todas las representaciones semánticas de los textos que incluyen el verbo *enviar* tendrían la valencia de “canal”, independientemente de su realización sintáctica en el texto.

Sin embargo al considerar la frase *Arturo envió a Víctor a traer las pizzas*, la segunda valencia ¿qué se envía? incluye ahora una persona, y en este caso, la cuarta valencia ¿vía que canal?, es dudosa. Por lo tanto se vuelve inconsistente la consideración inicial de valencias.

Por lo que se consideran, inicialmente, dos posibilidades para resolver esa inconsistencia:

Dividir el verbo *enviar* en dos homónimos, *enviar*$_1$ y *enviar*$_2$, el primero sería para los objetos de información y el segundo para personas, con un sentido relacionado a “mandar” o “dar órdenes”. Estos dos homónimos tendrían diferente número de valencias. El verbo *enviar*$_1$ tendría la cuarta valencia, y *enviar*$_2$ sólo tres valencias.

Dejar el verbo como una sola entidad, con sólo tres valencias, sin reflejar el canal en la definición lexicográfica. Entonces, toda la información no considerada es adicional y se trataría como circunstancial.

A continuación presentamos los patrones de manejo correspondientes. En el primer caso serían dos homónimos:

enviar$_1$

Person X send thing Y to entity Z through medium W

1 = X; *quién?*

1.1 N(an) %Víctor

2 = Y!; ***qué?***

2.1 N %~un paquete

3 = Z; a quién?

3.1 *a* N(an) % a Arturo

4 = W; vía qué?

4.1 por N(an) % por mensajería internacional

enviar$_2$

Person X send person or thing Y to make (if Y=person) or receive (if Y = thing) action Z.

1 = X; ***quién***?

1.1 N(an) %Víctor

2 = Y!; qué? o a quién?

2.1 N %~sus camisas

2.2 *a* N(an) %~a Arturo

3 = Z; ***a qué***?

3.1 *a* V_INF %~ a bañar

%~ a planchar

Se observa que este último patrón podría separarse, a su vez, en dos homónimos *enviar*$_2$ y *enviar*$_3$, para separar la disyuntiva de hacer o recibir la acción. De esta forma, *enviar*$_2$ correspondería a la valencia Y para una persona, y *enviar*$_3$ corresponde a la valencia Y para cosas.

En el segundo caso se consideró un solo patrón:

enviar

Person X send thing or person Y to entity or action Z

1 = X; *quién?*

1.1 N(an) %Víctor~

2 = Y!; qué o quién?

2.1 N %~unas camisas / ~unas cartas

2.2 *a* N(an) %~a Arturo

3 = Z; a quién? o a qué?

3.1 *a* N(an) % a Arturo

3.2 *a* V_INF % a planchar

En el enfoque de constituyentes hay un desarrollo particular para este caso. En la GB se consideró que todo verbo debe tener un sujeto aún si se trata de la forma infinitiva. En la frase *Víctor mandó a Arturo a traer unas pizzas* y siguiendo a la GB, *Arturo* es el sujeto de *traer* y al mismo tiempo objeto directo de *enviar*. En la frase *Arturo mandó sus camisas a planchar* y siguiendo nuevamente a la GB, [Lamiroy, 94] considera que ambos verbos comparten el objeto directo: *sus camisas*.

En la gramática española, [Seco, 72] considera que cuando los verbos aparecen en infinitivo, no expresan por sí mismos el tiempo en que ocurre la acción sino que se deduce del tiempo de la oración, y en cuanto a los sujetos, las frases con verbos en infinitivo presentan los siguientes casos:

- El sujeto es indeterminado. Ejemplo: *querer es poder*.
- Es sujeto (infinito sustantivizado). Ejemplo: *el murmurar de las fuentes, el comer bien*.
- El sujeto es el mismo sujeto del verbo principal. Ejemplo: *pelaremos hasta morir, deseo pasar unas vacaciones muy tranquilas*.
- El sujeto del infinitivo es distinto del sujeto del verbo principal. Ejemplo*: Por no saber yo nada me sorprendieron. Te prohíbo hablar.*

Por otra parte, [Gili, 61] indica que cuando el infinitivo es complemento directo se construye sin preposición, por ejemplo *oigo tocar las cornetas*. Con verbos de mandato no hay dificultad porque el infinitivo es la cosa mandada y su sujeto es un claro complemento indirecto; pero con verbos de percepción la cuestión resulta difícil. El autor considera que mirada la cuestión psicológicamente, el infinitivo y su sujeto forman una representación conjunta que actúa en su totalidad como complemento directo del verbo principal. Cuando el infinitivo es complemento indirecto lleva preposición. Realmente, en los textos, las valencias semánticas relevantes encuentran su expresión explícita en las palabras correspondientes conectadas (dependientes sintácticamente) con la palabra predicativa, directamente o a través de preposiciones.

Estos estudios muestran los diferentes aspectos que se deben considerar con los verbos en infinitivo; su construcción sintáctica y su información léxica. Apoyan la descripción individual de las variaciones sintácticas del verbo específico para el análisis sintáctico así como el análisis de la construcción sintáctica en la que aparecen. En los patrones de manejo, se determinan las valencias semánticas y se describe esas posibilidades de realización con infinitivo. En los ejemplos anteriores, *planchar* y *traer* son acciones, ordenadas por el sujeto de *enviar* y ejecutadas o recibidas por el tipo de objeto directo que emplea el verbo *enviar*.

Así que la MTT a través de los patrones de manejo nos permite escoger para palabras predicativas cualquier número de valencias, pero con reservas acerca de dominios semánticos restringidos y con advertencias acerca de posibles efectos y contradicciones cuando se procesan textos sin restricciones. Las implicaciones lexicográficas se detallan en la sección 2.10

2.10 MÉTODOS TRADICIONALES PARA CARACTERIZAR FORMALMENTE LAS VALENCIAS

Subcategorización

Los métodos tradicionales más empleados para describir el nivel sintáctico de los lenguajes naturales son los basados en las gramáticas generativas, y el español no es una excepción. En esta aproximación, la descripción de las características de las valencias para los lexemas se realiza, principalmente para los verbos. Los actuantes se describen en ellas desde un punto de vista puramente sintáctico (denominados complementos). Las valencias se denominan características sintácticas. La información de la estructura de los complementos de un verbo se conoce como subcategorización. Cada subcategoría del verbo tiene su propio conjunto de complementos que usualmente van en un orden lineal predeterminado.

A continuación presentamos ejemplos de subcategorización para el español, para una selección de verbos.

- *ver* tiene una subcategoría *grupo nominal* (GN), por ejemplo: *Beto vio una araña.*
- *dar* tiene la subcategoría GN seguido de GP, por ejemplo *Beto dio una carta a su novia* y la permutación GP GN, por ejemplo *Beto le dio a su novia una carta*.
- *poner* también tiene la subcategoría GN seguido de GP, el grupo preposicional se realiza mediante diferentes preposiciones: *en, sobre, bajo,* etc. Por ejemplo: *Beto puso los libros en el librero*. En este ejemplo el grupo preposicional es introducido por la preposición *en*. Notamos que aunque se especifique esta construcción (*en* GN) existen otros grupos

preposicionales con la misma preposición como: *en la mañana*, que no corresponde a la realización de la misma valencia. Con la aproximación de subcategoría se pierde esta información para el nivel semántico.

- *Llover* no tiene subcategoría, puesto que es intransitivo y no permite complementos, se ve natural para un verbo impersonal.
- *acusar* tiene la subcategoría GN *de*_INF (entre otras), y el objeto directo está conectado mediante la preposición *a*. Por ejemplo *Beto acusó a sus compañeros de negar su ayuda*.

Por supuesto, muchos verbos pueden asignar varias subcategorías. Por ejemplo, el verbo *decir* asigna: GN, *que*_O, *a* GN seguido de GN. Por ejemplo: *dijo unas palabras de aliento, dijo que el director vendrá pronto, dijo a sus compañeros unas palabras de aliento*. Desde el punto de vista de la subcategorización, también existen otros verbos cuyas estructuras de complementos en algunas oraciones son similares a esta última subcategoría. Por ejemplo: *aconsejó a su alumna* (*a* GN) *una vez más* (GN), ya que no se distingue que el último grupo nominal representa una circunstancia, no directamente relacionada con el significado del verbo.

Los marcos de subcategorización se emplean desde hace mucho tiempo [Boguraev *et al*, 87], ya que son útiles para restringir el número de análisis generados por el analizador sintáctico, para la generación automática de texto y para aprendizaje de lenguajes. Debido a esta utilidad muchos esfuerzos manuales se han aplicado a su compilación para tareas de procesamiento lingüístico de textos por computadora, principalmente para el inglés: ALVEY [Boguraev *et al*, 87] y COMLEX [Grishman et al, 94].

La subcategorización se describe en diccionarios modernos como COMLEX [Grishman *et al*, 94], mediante la descripción de los constituyentes que lo forman, principalmente, y algunas otras características. Para el desarrollo de este diccionario emplearon las clasificaciones verbales de diversos diccionarios. Para ilustrar la estructura de las entradas presentamos tres ejemplos:

```
(verbo :orth “aceptar”    :subc ((gn)
                                 (que-o )
                                 (gn)))
(sust  :orth “aceptación”)
(verbo :orth “abstenerse”      :subc ((intr)
                                 (gp         :val (“de”))
                                 ))
```

El primer símbolo (verbo, sust) marca la categoría gramatical o POS. La

característica *orth* describe la forma de la palabra. Las palabras para las cuales se consideran sus complementos tienen la característica *subc*. Por ejemplo, para el verbo *abstenerse* se definen dos tipos de subcategorización, el nombre (intr, pp) corresponde al nombre del marco. Se observa la consideración de que algunos verbos pueden pertenecer a más de un tipo de subcategorización.

Cada tipo de complemento se define formalmente mediante un marco. Cada complemento se designa por los nombres de sus constituyentes, junto con unas pocas marcas para indicar casos especiales como el fenómeno de control. El marco incluye la estructura de constituyentes (*cs*), la estructura gramatical (*gs*), una o más características (*features*) y uno o más ejemplos (*ex*). Por ejemplo:

(vp-frame s :cs ((o 2 : que-comp opcional))

:gs (:sujeto 1 : comp 2)

: ex “ellos aceptaron (que) era demasiado tarde”)

Donde los elementos de la estructura de constituyentes están indexados. En los campos de la estructura gramatical se indican estos índices, por ejemplo, el índice “1” se refiere al sujeto superficial del verbo. La “o” significa que es de tipo oración. Entre las características que se pueden definir y que en este ejemplo no están presentes, se encuentran: sujeto de ascensión, sujeto control, etc.

De los ejemplos anteriores se ve que en los marcos de subcategorización, generalmente, el orden de los complementos es fijo y todos los complementos aparecen después del verbo. Por ejemplo, para el verbo *abandonar,* un marco de subcategorización es un grupo nominal seguido de una frase preposicional introducida por la preposición *a*, es decir, NP GP(*a*). La permutación GP(*a*) NP puede existir, solamente si se expresa explícitamente con otro marco. Esta descripción es muy útil en inglés por su orden de palabras más estricto. En el español, este orden es más libre, por ejemplo, la frase *expresó*(V) *sus ideas* (NP) *con palabras sencillas*(GP) puede expresarse de diferentes maneras, por ejemplo: *expresó*(V) *con palabras sencillas*(GP) *sus ideas*(NP) o *con palabras sencillas*(GP) *expresó*(V) *sus ideas* (NP) son igualmente posibles y esas permutaciones son muy usuales.

Como presentamos en el capítulo primero, la información de subcategorización ha sido considerada en la mayoría de los formalismos gramaticales modernos. Inclusive se han llevado a cabo esfuerzos para estandarizar la información de subcategorización, principalmente por [EAGLES, 96], pero realmente la información de subcategorización en los diccionarios prácticos se ha definido considerando el aspecto teórico del formalismo considerado o las necesidades requeridas en la aplicación para la cual fueron construidos, o ambos.

Los diccionarios prácticos para el procesamiento lingüístico de textos por computadora pueden ser más prescriptivos o menos, dependiendo de sus bases teóricas (formalismos en los que se basan) o del propósito de aplicación. Por ejemplo,

considerando los diccionarios ILCLEX [Vanocchi et al, 94], ACQUILEX [Sanfilippo, 93], COMLEX [Grishman et al, 94] y LDOCE [Procter, 87] se observa lo siguiente:

1. El número de argumentos sólo se codifica explícitamente en ILCLEX (mediante una característica con valor numérico), en los demás se debe inferir.
2. La categoría sintáctica se indica en todos los diccionarios explícitamente, salvo en ACQUILEX. Éste sigue el formalismo de gramáticas categoriales, que especifica categorías simples y complejas, por lo que la categoría sintáctica se infiere.
3. Todos especifican requerimientos léxicos, por ejemplo la selección de una preposición particular para introducir complementos, aunque lo hacen con diferentes grados de granularidad.
4. La variación de marcos aparece explícitamente en todos, salvo en LDOCE, dónde se infiere. Pero varían considerablemente en la forma de codificarla y en el rango en que consideran este fenómeno. Por ejemplo, la opcionalidad de argumentos no siempre se trata como variación.
5. La estructura de roles semánticos sólo se marca en ACQUILEX, siguiendo la expresión de relaciones temáticas de [Dowty, 91].

Usualmente en los marcos de subcategorización el orden de los complementos es fijo y todos los complementos aparecen después del verbo. Esta descripción es especialmente útil para el inglés por su orden de palabras más estricto. En español el orden de palabras es más libre, aunque no totalmente, para lenguajes con un orden libre se estudian otras descripciones [Rambow & Joshi, 92], [Bozsahin, 98]. Aún cuando [EAGLES, 96] considera varias lenguas europeas (el español entre ellas), en su trabajo de recomendaciones de normalización, no consideran fundamental la información del orden lineal de los complementos. Sin embargo, explícitamente dicen que en algunas lenguas las restricciones en la precedencia lineal pueden ser completamente necesarias.

Presentamos los marcos de subcategorización para el verbo *acusar*. El ejemplo considera las ocurrencias con más del 10% en un corpus.

1. *a* GN
2. *de* GN
3. *de* V_INF
4. *a* GN *de* GN
5. *a* GN *de* V_NF
6. *Ø*
7. GN

De los siete marcos presentados, cinco corresponden al verbo *acusar*$_1$, el sexto

marco corresponde a oraciones que presentan antes del verbo el complemento directo, y el séptimo marco corresponde al verbo *acusar*$_2$, con sentido de "poner de manifiesto" o "revelar". Este ejemplo muestra la cantidad de clases de subcategorías consideradas en esta aproximación, y su generalidad que origina la eliminación de información relevante.

Patrones de manejo

La estructura que permite la asociación de las valencias semánticas y sintácticas es el patrón de manejo de un lexema. Este PM es pues una noción lingüística muy importante que se describe con más detalle a continuación. Los patrones de manejo sintáctico constan de cuatro secciones:

Primera sección

La palabra encabezado, que corresponde al verbo considerado con un significado específico. Para diferenciar los patrones de manejo sintáctico de verbos homónimos, se da una numeración, por ejemplo: **alternar1** (tener trato con otras personas) y **alternar2** (hacer dos o más acciones una tras otra y repetidamente). Para diferenciarlos, la numeración es totalmente arbitraria pero debe existir al menos un elemento diferente en el patrón de manejo sintáctico respecto de los otros.

Segunda sección

La explicación semántica de la situación relacionada a cada verbo específico. En los ejemplos que mostramos, optamos por una simplificación del método de descripción del modelo en la MTT, la explicación semántica se reemplaza por una oración simple en inglés.

En esta sección se definen las valencias, cuyo orden es hasta cierto grado arbitrario, aunque cada lexema normalmente impone un cierto orden "natural" en las valencias, indicando primero los más importantes. Este orden a veces concuerda con el orden en la oblicuidad. Por ejemplo, en primer lugar una entidad activa, el sujeto, enseguida el objeto principal de la acción (primer complemento), después otros complementos, si existen.

También la forma sintáctica de expresar las valencias influye significativamente en el orden. Por ejemplo, cuando el objeto directo se conecta directamente a la palabra encabezado, sin preposiciones, va antes del complemento indirecto el cuál se conecta generalmente mediante preposiciones.

Para cada valencia sintáctica se indica la valencia semántica correspondiente. En el ejemplo presentado en la sección 2.2 la fórmula 2 = Y indica que la valencia sintáctica 2 corresponde con la valencia semántica Y. Generalmente, el orden de las valencias sintácticas y semánticas es el mismo.

Tercera sección

La descripción de cada valencia sintáctica. La lista exhaustiva de todas las posibles formas de realización de cada valencia sintáctica, en los textos. Se numeran para cada valencia, para la n-ésima, serán *n1, n2, ... nk*, donde *k* depende del lexema específico y de la valencia. El orden es arbitrario aunque se prefiere que aparezcan primero las formas más frecuentes. Todas las posibles opciones se expresan con símbolos de categorías gramaticales o subclases de lexemas muy específicas, por ejemplo: S para sustantivos, V para verbos, Adj para adjetivos, etc. También se especifican las palabras específicas que aparecen antes de estos símbolos, como las preposiciones o conjunciones (*que*), en la forma literal en que se encuentran en los textos.

Después de las categorías gramaticales pueden seguir parámetros léxicos relevantes para el nivel sintáctico. Por ejemplo, la categoría *an* indica que este actuante es una entidad animada y el parámetro *na* que indicaría lo opuesto, entidad no animada. La marca INF que indica infinitivo para los verbos, etc. También pueden seguir, a las categorías gramaticales, los descriptores semánticos relevantes para el nivel sintáctico. Por ejemplo *loc* que indicaría locativo.

Por último, esta sección contiene el indicador de su condición de obligatoriedad en los textos. Si no está presente este indicador, significa que es opcional en su realización sintáctica, es decir, que aunque semánticamente existe, se desconoce el actuante, es algo o alguien no especificado. El reconocimiento de estos actuantes se lleva a cabo en niveles más profundos del análisis.

Cuarta sección

En la última sección se muestra la información acerca de los posibles ordenamientos o combinaciones de valencias sintácticas, es decir, de los órdenes posibles e imposibles. Para lenguajes con un orden de palabras con menos restricciones, esta lista pudiera reducirse a la lista de órdenes imposibles. Por ejemplo:

- ÓRDENES POSIBLES 2.1, 3.2, 4.1

  significa que la primera opción de la segunda valencia seguida de la segunda opción de la tercera valencia seguida de la primera opción de la cuarta valencia aparece en los textos.

- ÓRDENES IMPOSIBLES 2.2, 3.1

  significa que la segunda opción de la segunda valencia seguida de la primera opción de la tercera valencia nunca parece en los textos.

Pudiera establecerse también que con la misma notación se describen todas las combinaciones de las opciones especificadas. En lenguajes con un orden de palabras estricto, las combinaciones posibles describen ese orden.

A continuación se presenta un ejemplo de los patrones de manejo sintáctico, para el verbo *solicitar*. Las preposiciones se marcan en tipo itálico, los ejemplos se encuentran después del signo **%**, el signo ~ se utiliza para colocar la palabra encabezado y en la sección de combinaciones, el "0" representa la palabra encabezado.

solicitar

X asks something Y from Z

Número	Patrón de manejo	Ejemplo
X = 1; who asks?		
1.1	S (an)	Juan / el gobierno ~
Y = 2; what?		
2.1	S (na)	~ una prórroga / un préstamo
2.2	que C	~ que este libro se le dé
Z = 3; from whom?		
3.1	*a* S (an)	~ a la secretaria
3.2	*con* S (an)	~ con el secretario
3.3	*de* S (an)	~ de usted
3.4	*en S* (na)	~ en urgencias
POSIBLE:		
(1) 0 2 3	(El partido) solicita una prórroga al gobierno.	
(1) 0 2	(Ella) solicita un préstamo.	
IMPOSIBLE:		
(1) 0	*(El partido) solicita.	
(1) 0 3	*(El partido) solicita al gobierno.	

donde:

S- sustantivo o pronombre personal

que C - cláusula subordinada relacionada a la principal a través de *que* (... *que este libro se dé al muchacho*)

(an)- animado (solamente para sustantivos), corresponden a criaturas vivientes, incluyendo al ser humano, grupos de humanos, organizaciones,

etc.

(na) - inanimado (solamente para sustantivos), como argumento, acción y lugar,

Existen otras abreviaturas que no se utilizaron en este ejemplo, como:

V - verbo

Adj - adjetivo

Adv - adverbio

Pp - pronombre personal

Q - cláusula subordinada que tiene forma de interrogación. Por ejemplo, para el verbo *decir*: "*Dijo: ¿A quién se dio este libro?*"

(inf) - forma infinitiva (solamente verbos)

(tm) - intervalo de tiempo (solamente para sustantivos)

(mn) - de manera (solamente para adverbios)

(pc) - de lugar (solamente para adverbios)

(nom) - caso nominativo (solamente para pronombres personales)

(acc) - caso acusativo (solamente para pronombres personales)

(dat) - caso dativo (solamente para pronombres personales)

(inc) - caso inclusivo (solamente para pronombres personales)

Este caso, *inclusivo* lo introducimos como una designación de las formas contraídas *conmigo, contigo, consigo*.

Se observa en este ejemplo, que las preposiciones, principalmente, son los lexemas de conexión que introducen los objetos del verbo. El uso de las preposiciones es muy variado en el español y los complementos de los verbos, generalmente, exigen el empleo de una determinada preposición. Por ejemplo: *me arrepiento de mis acciones, lo expresó con ademanes, insisto en pagar*. Esto ocurre también con sustantivos y adjetivos que exigen el empleo de una determinada preposición. Ejemplos: *intolerante con sus amigos, esencial en el proyecto, inferior a su compañero*. En cuanto al objeto directo, normalmente se construye sin preposición salvo cuando designa seres humanos o animados que podrían aparecer en la posición de sujeto.

Existen pues varias diferencias importantes de la aproximación de subcategorización respecto a las gramáticas de dependencias:

- Se postula un marco, como un conjunto de subcategorías y después se intenta clasificar la diversidad total de verbos para ese marco. Esta aproximación es suficientemente buena cuando el número total de

subcategorías es pequeño, pero no así en lenguajes donde casi cada verbo presenta su propia subcategoría específica, como en español.

- Generalmente, no intentan establecer correspondencia entre valencias sintácticas y semánticas. La separación entre complementos del verbo y complementos circunstanciales no existe, por lo que pueden incluirse predicados cuya ocurrencia es obligatoria en el contexto local de la frase pero que no son seleccionados semánticamente por el verbo.
- Usualmente, en cada subcategoría, las valencias sintácticas se consideran en un orden fijo predeterminado. Por ejemplo, los complementos preposicionales en una frase como (*persona A*) (*expresa*) (*idea B*) (*mediante C*), corresponde exactamente a una regla de producción dando los constituyentes justamente en el mismo orden fijo: *GN GV GN GP*. Si se añaden complementos circunstanciales como *en la mañana* o se emplea una variación sintáctica como en la frase *Javier expresa mediante tiernas palabras sus sentimientos* estas reglas fallarán y será necesario incluir nuevas reglas.

Los marcos de subcategorización consideran un conjunto de complementos. La colección de esos marcos para lenguajes como el inglés no es vasta. En español, como en otros lenguajes, la variedad en el uso de preposiciones es tan amplia que la colección completa sería muy grande y muchas clases de subcategorización se requerirían para describir un verbo.

2.11 LOS PATRONES DE MANEJO AVANZADOS, COMO UN MÉTODO ALTERNATIVO

Como se expuso en la sección 1.2 y en la sección anterior, las descripciones de la estructura superficial en la MTT están orientadas a los seres humanos. Aunque toda la información de los PM es necesaria, no debemos imponer la estructura del formalismo ya que para nosotros la finalidad de su uso es el procesamiento lingüístico de textos por computadora. Además, la estructura de los patrones de manejo debe modificarse para ayudar a identificar, clarificar y comparar las piezas de su información, con la finalidad de facilitar el diseño de un diccionario de PM, su uso y reuso. Cuando se construye un diccionario, uno de los objetivos es la generalidad del formato, y la posibilidad de una organización de trabajo modular.

En la nueva estructura formal que proponemos, considerando las caracterizaciones del español expuestas en las secciones anteriores, además de modernizar su formato, nos basamos en los sistemas de análisis sintáctico que dan mayor importancia a los diccionarios, donde cada característica se representa dando su nombre y sus valores, con múltiples valores permitidos para cada palabra. La capacidad que hace posible almacenar generalizaciones sintácticas en el diccionario es el sistema de pares de atributo valor, por ejemplo [Pirelli *et al*, 94], [Flickinger, *et al*, 85], [Santorini, 90], [Marcus *et al*, 94].

El sistema de atributo – valor se ha empleado en varios formalismos, especialmente se observa en el formalismo HPSG que utiliza las denominadas matrices de atributo – valor (AVM en inglés). Ejemplos de varias palabras, con esta descripción mediante AVM, se presentaron en la sección 1.2.

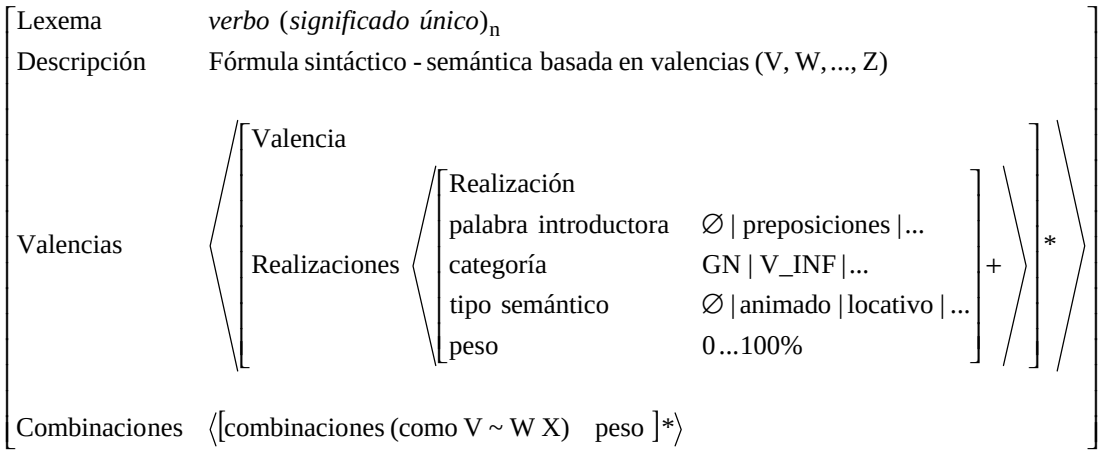


Figura 15 Patrones de manejo avanzados

Aquí: + denota uno o más elementos; * denota cero o más elementos; ~ denota el verbo.

La nueva estructura la denominamos patrones de manejo avanzados (PMA), en primer lugar para evitar el nombre ligado especialmente a la MTT y en segundo para tener un formato orientado a las computadoras.

La información contenida en los PAM corresponde a la expuesta en el capítulo anterior, y a la considerada en el formalismo de la MTT (en la tabla de PM), salvo la indicación de obligatoriedad de la presencia de cada valencia. En un PMA, la indicación de obligatoriedad, las posibles combinaciones de actuantes y las combinaciones prohibidas las hemos considerado de otra forma.

En la **Figura 15** presentamos la estructura formal y la notación de los PMA. El primer atributo denominado *Lexema* corresponde a la primera sección de los PM, la palabra encabezado. Su valor corresponde al lexema numerado con un sentido específico y una realización sintáctica particular, por ejemplo ***querer*$_2$**

El segundo atributo, denominado descripción, corresponde a la segunda sección de los PM, la explicación semántica de la situación relacionada a cada verbo específico[30], por ejemplo: *person X desires thing Y*

El tercer atributo, corresponde a la tercera sección de los PM, la tabla de patrones de manejo, donde las realizaciones de las valencias sintácticas se describen

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

recursivamente con una matriz atributo – valor. En cada realización se permiten los siguientes atributos:

- Palabra introductora
- Categoría gramatical
- Tipo semántico
- Peso

Las palabras introductoras son, principalmente, preposiciones simples o complejas, aunque también pueden ser palabras que introducen cláusulas subordinadas, como *que,* o en el caso de una realización directa con grupo nominal, realmente no está presente. Las categorías gramaticales son de cualquier tipo.

Los descriptores semánticos pueden ser de diversos tipos, nosotros hemos considerado principalmente la animidad y la locatividad. La primera en forma detallada en la sección 2.4, la segunda en un ejemplo de la sección 2.2. La consideración de descriptores semánticos, como locatividad, se ha considerado en trabajos recientes. [Bleam *et al*, 98] llegan a la conclusión de que para capturar propiedades léxico semánticas, que ayuden a reducir las variantes en el análisis sintáctico, es necesario introducir características de propiedades semánticas. La diferencia con su trabajo es que ellos definen una clase de preposiciones locativas (no específicas para cada verbo dado) e imponen una restricción en un nodo del árbol elemental[31] para los verbos de movimiento que utilizan esa misma clase. Un punto importante de convergencia es que consideran la necesidad de separar las frases preposicionales cuyo significado está implícito en el verbo, de las demás.

El peso considerado en las realizaciones define las probabilidades de llenado de diferentes valencias. Por ejemplo, en las frases siguientes, la segunda valencia del verbo *acusar* aparece realizada de tres formas diferentes: como *a* GN, como pronombre reflexivo y como clítico.

A quienes acusan de comportamiento arrogante.

El fiscal me acusa de delito de alta traición.

Acusándole de ser el sostenedor y portavoz de Mario Segni.

Y cada una de ellas tiene una probabilidad diferente. En los ejemplos siguientes, la tercera valencia del verbo *solicitar* aparece realizada con diferentes preposiciones introductoras:

Solicitará al seleccionador argentino Alfio Basile la posibilidad de volver a jugar con Argentina.

El Consejo Superior de Deportes solicita de la Subsecretaría del Ministerio

[31] Los árboles elementales en las “Gramáticas de adjunción de árboles” representan un cierto tipo de subcategorización para una clase de verbos.

de Cultura la designación de dos inspectores técnicos.

Los aficionados solicitaron unos pases con el delegado.

Y de entre estas realizaciones algunas son más frecuentes que otras, es decir, tienen diferentes probabilidades. La obligatoriedad queda implícita en este peso. Si una valencia tiene presencia en todas las oraciones extraídas del corpus para un verbo específico, se considera como una evidencia de obligatoriedad.

El último atributo, corresponde a la cuarta sección de los PM, los ejemplos de combinaciones posibles y de las combinaciones no permitidas. Entre las dificultades que se presentan para definir los ejemplos de esta sección se encuentran los siguientes:

- No deben ser aleatorios.
- Se basan en experiencia.
- Se requiere que sean exhaustivos.

Lo que implica que los ejemplos posibles e imposibles se deben describir por personas muy calificadas. Además de esto hay que considerar que el español tiene un orden de palabras más libre que el inglés pero no totalmente libre, por lo que las posibles combinaciones de valencias son limitadas. A partir de la indicación de obligatoriedad se pueden definir algunas combinaciones no deseadas, pero no la totalidad. Las combinaciones posibles y las prohibidas pueden definirse basándose en cierta experiencia pero no reflejarían los cambios en el lenguaje ni las preferencias en dominios específicos. Por lo que para adquirir esta información consideramos la obtención de pesos estadísticos.

Para el inglés funciona bien buscar usualmente todos los objetos del verbo después de él. Sin embargo, para el español, la información de posibles posiciones de la valencia es necesaria para el analizador sintáctico. Por ejemplo, en las frases 1, 2 y 3, anteriores, el objeto indirecto no aparece después del verbo, de tres maneras distintas: 1) en la forma *a* GN antes del verbo, 2) como pronombre reflexivo entre sujeto y verbo, y 3) como clítico dentro del verbo.

Así que además de la información determinística, incluimos en los PAM información de evaluación, en forma numérica de probabilidades de diferentes opciones. La información de evaluación incluye:

- Probabilidades de llenado de diferentes valencias.
- Probabilidades del uso de diferentes opciones de la misma valencia.
- Medidas de compatibilidad de varias combinaciones de opciones específicas para diferentes valencias.

Esta información, determinística y probabilística, es muy útil para el procesamiento lingüístico de textos por computadora. Esta información tiene uso

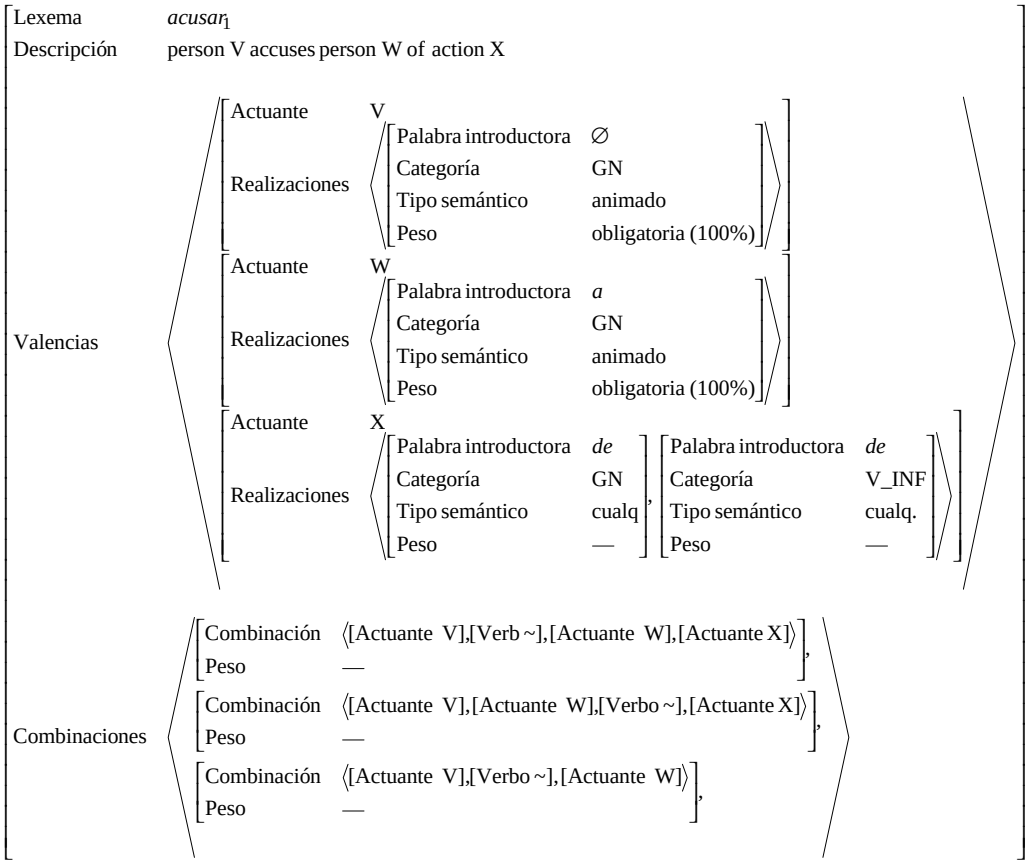


Figura 16. Estructura formal para el verbo *acusar*

inmediato en el análisis sintáctico, y en filtros para rechazar resultados intermedios imposibles o no deseados. Por ejemplo, el analizador sintáctico empleará esta evidencia para buscar las valencias aún en enlaces distantes. Si el verbo *acusar* requiere forzosamente la presencia del objeto directo, con esta indicación, el analizador sintáctico buscará este pedazo de información alrededor del verbo, considerando también las probabilidades de su aparición antes y después del verbo.

La obtención de datos estadísticos confiables para evaluación estadística es muy difícil y aún algunas veces el uso de estimaciones subjetivas previas (inventadas por el investigador) es mejor que la ignorancia total de esa información. Entonces, para compilar los PAM se requiere información sintáctica, estadística y conectada con la semántica. En la parte semántica es necesario incluir la marca de animidad y de locatividad en el corpus. Además, se requiere detectar la llamada atracción léxica (ocurrencia concurrente en estructura sintáctica) entre los verbos y las preposiciones que introducen las valencias y diferenciar las valencias correspondientes a diversos significados del verbo.

En la **Figura 16** presentamos el PAM del verbo *acusar*. Ya que no existen diccionarios para el español con información completa de subcategorización, consideramos la información de varios autores. Por ejemplo [Penadés, 94] considera

el verbo *acusar* entre 145 verbos que analizó, con el siguiente esquema sintáctico-semántico:

Alguien	acusa	a alguien	de algo
agente puro causativo interno directo	Acción causativa intrínseca directa	afectado especificado	especificación

Entre otros autores, [Alonso, 60] muestra algunos ejemplos de empleo: esquema sintáctico-semántico: *a alguno al*, *ante el juez*, *de haber robado*, *de los pecados* (verbo reflexivo), *de lo mal que se ha portado* (verbo reflexivo). [Nañez, 95] presenta el uso de preposiciones en orden alfabético para construcciones sintácticas; para el autor, el verbo *acusar* emplea las preposiciones *a*, *ante*, *de*; también muestra algunos ejemplos de uso en la misma forma que Alonso.

Con esta información no es posible llenar completamente los PAM y aún alguna información considerada requerirá comprobación con bases de datos de textos o con la experiencia de recursos humanos calificados. Los campos de los pesos quedan con la marca — que indica ausencia de datos. En el capítulo cuatro presentamos la adquisición de estos pesos mediante un método automático.

CAPÍTULO 3. ANÁLISIS SINTÁCTICO Y DESAMBIGUACIÓN BASADA EN PATRONES DE MANEJO AVANZADOS

En este capítulo presentamos el modelo general que proponemos para el análisis sintáctico del español y la desambiguación. Primero describimos el modelo general y posteriormente los elementos del modelo. Comenzamos con la gramática generativa, la transformación de los árboles de constituyentes a los árboles de dependencias, y su algoritmo de análisis sintáctico. Posteriormente presentamos el algoritmo de proximidad semántica y su aplicación a la desambiguación sintáctica. Al final detallamos la asignación cuantitativa a las variantes y las posibilidades del modelo de votación para la desambiguación sintáctica.

3.1 ANTECEDENTES DEL SISTEMA PROPUESTO

[Briscoe, 96] afirma que a pesar de más de tres décadas de investigación no ha sido posible desarrollar un analizador sintáctico práctico, independiente del dominio, de textos sin restricciones. El autor considera que para obtener esa clase de analizador sintáctico, que dé por resultado un análisis correcto o un análisis útil aproximado en el 90% de las oraciones de entrada, es necesario solucionar al menos los tres problemas que crean dificultades severas en los analizadores sintácticos convencionales que emplean algoritmos de análisis con una gramática generativa: delimitación de grupos sintácticos debido a elementos de puntuación[32], desambiguación por la gran cantidad generada de variantes de estructuras, y la insuficiencia de cobertura.

Estos tres aspectos que puntualiza Briscoe son muy importantes y presentan algunas características específicas en cada lenguaje, además de interdependencias entre ellos, que a continuación presentamos:

1) El problema de la delimitación de grupos sintácticos se ha intentado solucionar introduciendo la puntuación a las reglas de la gramática [Jones, 94], [Osborne, 96]. En lenguajes donde existen reglas claras para una puntuación estricta, la inclusión de reglas de puntuación en las reglas de la gramática generativa ayuda a eliminar variantes. En cambio, en los lenguajes donde la puntuación no se define de manera estricta, como es el caso del español, la inclusión de condiciones de puntuación ocasiona el aumento de la cantidad de reglas de la gramática. Este hecho también incide en la disminución de cobertura, por la imposibilidad de definir todas las posibilidades de puntuación de textos arbitrarios.

[32] Ejemplos de este problema son las oraciones que contienen textos adjuntos delimitados por guiones, paréntesis o comas que no siempre se encuentran en una relación sintáctica con el texto circundante.

El empleo de procesos de edición previos al análisis, para delimitar los constituyentes haría menos complejo el análisis sintáctico. Sin embargo, esta tarea requeriría reglas claras del uso de la puntuación en el lenguaje. Esto sin considerar otras características como la delimitación estilística, mediante comillas, guiones, apóstrofos, etc. la cual tiene una variedad mayor.

2) La insuficiencia de cobertura, es decir, tratar con casos de oraciones de entrada que están fuera de la cobertura sintáctica del sistema de reglas se ha considerado como un problema de labor intensiva y de compilación de cantidades extensas de conocimiento lingüístico, dada la propiedad de los lenguajes naturales de ser infinitos. Sin embargo, esa labor se tiene que detener en algún momento, por su imposibilidad de ser total. Esto debido a que cualquier modelo es limitado, no tiene una cobertura total del fenómeno que intenta representar. En el caso de las gramáticas generativas, cada una tiene su propia cobertura, siempre restringida.

La ampliación de la cobertura no se logra simplemente añadiendo más reglas, es necesario estudiar cómo afecta cada inserción a la gramática global. Además, como explicaremos más adelante, la cobertura se ve afectada por el grado de acierto de la gramática.

3) La desambiguación se requiere para disminuir la gran cantidad generada de variantes de estructuras. A mayor cobertura, menor número de restricciones y por lo tanto mayor cantidad de variantes. La introducción de mayor cantidad de reglas para la delimitación de constituyentes (por la falta de reglas precisas), también introduce otras posibilidades de enlaces de constituyentes y una cantidad adicional de variantes. Por lo que el problema a enfocar es la desambiguación.

Modelos empleados

Los modelos matemáticos del lenguaje [Uszkoreit, 96] son, básicamente, de dos tipos: los solamente simbólicos y los que adicionalmente aplican métodos estadísticos. Los simbólicos son sistemas formales axiomáticos compuestos por un conjunto de símbolos y de reglas, que establecen las combinaciones de símbolos. Se postulan propiedades generales sobre los símbolos así como sus relaciones, y a partir de estos axiomas se obtienen nuevas propiedades de manera deductiva. Ejemplos de estos modelos son los ya vistos en los enfoques de constituyentes y de dependencias.

Los modelos estadísticos fueron desarrollados a partir de la Teoría de la Información [Shannon, 49] y la estadística. Estos modelos describen el lenguaje como un conjunto de sucesos que presentan una determinada frecuencia; cada morfema, cada categoría sintáctica, cada sintagma, cada significado tienen una cierta probabilidad de aparecer en un determinado contexto. Los modelos estadísticos se fundamentan en los datos obtenidos a partir de corpus lingüísticos. La principal desventaja de los métodos estadísticos es que requieren una base estable, requieren corpus de textos que cuenten con todas las palabras necesarias y con frecuencias que

permitan su estudio, es decir, son métodos que requieren una base más objetiva. Con estos modelos no es posible distinguir si un grupo nominal es un objeto directo o si una frase preposicional es un objeto indirecto, tal vez sólo con corpus de tamaño de cientos de millones de palabras [Yuret, 98].

Estos modelos estadísticos, que aparecieron en los años cincuentas y sesentas y que fueron muy criticados, han recuperado el interés [Church & Mercer, 93] gracias al desarrollo tecnológico que permite tratar enormes cantidades de datos mediante computadoras y programas accesibles a los investigadores. Este nuevo auge también se debe al estancamiento de los resultados obtenidos con los métodos clásicos simbólicos [Charniak, 93]. Los modelos matemáticos, en distintas variantes, son los que se han empleado en las últimas décadas para realizar el análisis sintáctico de textos por computadora. El modelo que nosotros proponemos, también pertenece a esta clasificación.

Los analizadores sintácticos que se han desarrollado para el análisis sintáctico de lenguajes naturales se han basado en un único formalismo gramatical. Casos de este tipo son: [Grinberg *et al*, 95] basándose en la LG (*Link Grammar*); [Gawron *et al*, 82], [Briscoe & Carroll, 93] basándose en CFG; [XTAG, 95] basándose en las TAG.

Con estos modelos se genera una gran cantidad de variantes de análisis para cada oración que se procesa. Dado que usualmente, las oraciones de los lenguajes naturales tienen varios análisis sintácticos posibles, el problema en la desambiguación sintáctica es escoger el o los posibles análisis correspondientes a la intención del autor. Para realizar esta elección en el análisis sintáctico, dada su complejidad, se han aplicado adicionalmente otros métodos. Principalmente, se ha intentado la desambiguación sintáctica mediante esos mismos formalismos, enriquecidos con estadísticas. Por ejemplo [Schabes, 92] y [Carroll & Weir, 97] asocian información de frecuencias al formalismo LTAG (*Lexicalized TAG*), para gramáticas CFG se han desarrollado versiones probabilísticas, las cuales han sido investigadas desde [Suppes, 70], también por [Lari & Young, 90], [Kupiec, 91], [Jelinek *et al*, 92], [Charniak, 93], [Manning & Carpenter, 97], [Mohri & Pereira, 98] entre otros; para HPSG, [Brew, 95] presenta la versión estocástica; en analizadores sintácticos orientados por los datos [Bonnema et al, 2000].

[Baker, 82] demostró que la re-estimación de Baum-Welch podía extenderse a CFG en Forma Normal de Chomsky (CNF). [Fujisaki *et al*, 89] demostraron que el algoritmo de Viterbi puede usarse en conjunto con el algoritmo CKY (que se describe en la sección 3.6) y una gramática CFG en forma CNF. Sin embargo, las gramáticas sin restricciones rápidamente se vuelven impracticables porque el número de parámetros que requieren estimación se vuelve muy grande, y esos algoritmos son polinomiales en la longitud de la entrada y en el número de parámetros libres.

Idea de combinación de métodos

En esta investigación, consideramos que la resolución de la ambigüedad sintáctica requiere un sistema compuesto de un conjunto de métodos. Es decir, se requiere desarrollar un conjunto de módulos basados en modelos de tipos diferentes de conocimiento, que analicen las oraciones, y de sus resultados tomar la decisión final de cuáles son las variantes aceptables en base a una *votación*. De esta forma, cada uno de los módulos dará una medida cuantitativa de la probabilidad de una u otra variante de estructura, y finalmente el sistema completo elegirá las variantes con los valores máximos de esas evaluaciones estadísticas.

La idea no es muy nueva. En otras áreas como en el marcaje de POS se ha empleado esta misma idea. En este marcaje existen métodos híbridos que combinan diferentes aproximaciones, por ejemplo el uso de recursos basados en estadísticas y en conocimiento lingüístico, como en [Tzoukerman *et al*, 94]. [Samuelson & Voutilainen, 97] presentan una discusión comparativa de marcadores de partes del habla basados en lingüística y en estadística. [Padró, 98] usa relajación, un algoritmo iterativo para realizar optimización de funciones basada en información local, que también permite el uso de restricciones con múltiples características provenientes de diversas fuentes.

En el análisis sintáctico, se ha intentado emplear diferentes modelos como base de un método solo. Por ejemplo, [Abney, 91] se basa en estudios sicolingüísticos de [Gee & Grosjean, 83] para proponer el análisis sintáctico superficial. [Gee & Grosjean, 83] enlazan duraciones de pausa en la lectura y esquematización de oraciones *ingenuas*, a grupos de texto, que de una manera muy general corresponden a la separación de una cadena de palabras después de cada núcleo-*h*. El análisis sintáctico superficial analiza partes de la oración. La oración se segmenta en partes no traslapadas, el análisis de estos segmentos es la base del análisis sintáctico total, que detecta los argumentos del verbo y pospone decisiones de enlaces de grupos preposicionales.

[Magerman, 95] basa el análisis sintáctico en métodos estadísticos que reemplacen las habilidades de toma de decisiones del ser humano con algoritmos de toma de decisión. Emplea algoritmos de clasificación de árboles de decisiones, que además de identificar características relevantes para cada decisión y decidir la selección basándose en esas características, asignan una distribución de probabilidades a las elecciones posibles.

Para nosotros, dado que no podemos reproducir las habilidades humanas para entender una oración, el análisis sintáctico y su desambiguación debe basarse en modelos de conocimiento diverso. La elección de estructura debe hacerse en términos cuantitativos, asignando pesos, o evaluaciones estadísticas, a cada una de las variantes de estructura sintáctica. La variante con el peso más grande se considera como la mejor, mientras mayor sea el peso más posibilidades tiene de ser la variante correcta.

Una ventaja es que el carácter cuantitativo de la estimación permite la combinación de diferentes métodos para el análisis y su desambiguación.

3.2 ESTRUCTURA GENERAL DEL ANALIZADOR

El mutimodelo de generación que proponemos incluye principalmente los patrones de manejo, las reglas ponderadas, y la proximidad semántica. En la **Figura 17** presentamos el esquema general del análisis sintáctico propuesto, con resolución de ambigüedad. Cada uno de los métodos considerados tiene varias salidas con distintos pesos, que en la figura se representan mediante líneas, ordenadas de mayor a menor. Las variantes de cada grupo, con mayores probabilidades, constituyen la entrada al módulo de votación, donde se seleccionan las más adecuadas. Hacemos notar en este esquema, que a futuro este mismo sistema puede incluir otros modelos.

Los tres modelos que consideramos son modelos matemáticos y requieren de la compilación de diccionarios: el conjunto de PMA, el conjunto de reglas ponderadas, y la red semántica. Los tres son fuentes de conocimiento muy diferente, son recursos léxicos diferentes, y todos son necesarios porque contribuyen con distintos puntos de vista, para el análisis automático de textos sin restricciones. En esta investigación desarrollamos un método semiautomático para la compilación del diccionario de PMA, construimos el modelo de reglas ponderadas, y presentamos las bases para el modelo de proximidad semántica.

Patrones de manejo

Este método se basa en conocimiento lingüístico que adquieren los hablantes nativos durante el aprendizaje de su lenguaje, por lo que se considera el método principal. 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, por lo que se consideraron los otros modelos. En algunos casos, los modelos de proximidad semántica y reglas ponderadas resolverán la ambigüedad.

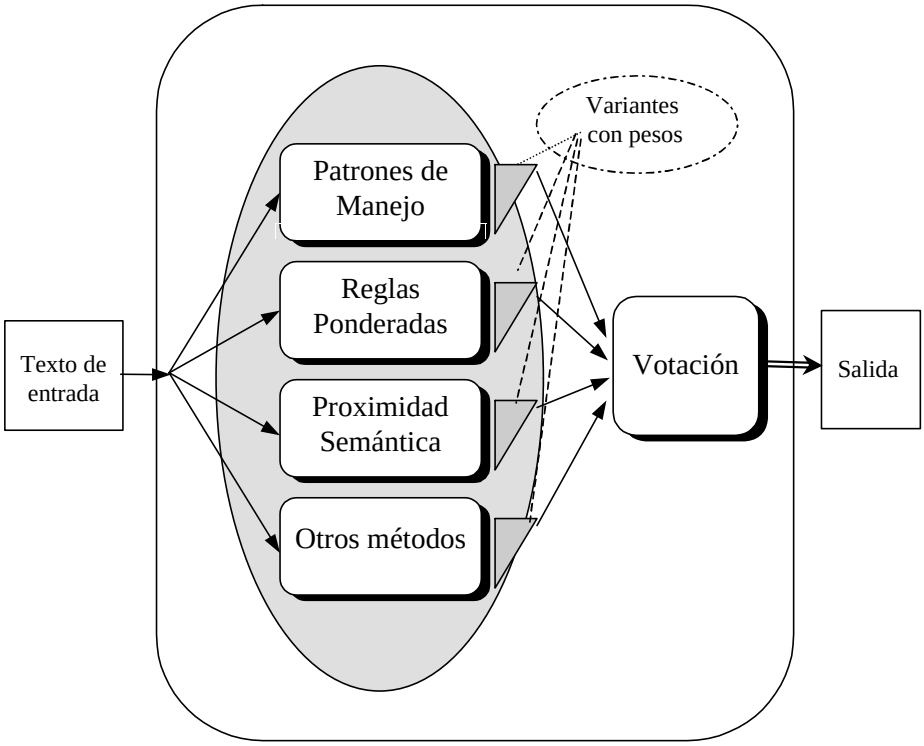


Figura 17. Estructura del analizador con resolución de ambigüedad

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.

En el capítulo anterior mostramos el análisis y desarrollo teórico de esta herramienta para el español. Para compilar el diccionario de PM desarrollamos un algoritmo iterativo y empleamos un corpus marcado con POS, este trabajo se detalla en el siguiente capítulo, dónde presentamos los resultados obtenidos.

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 gramáticas generativas que ya describimos en el primer capítulo. Se codifica directamente el conocimiento gramatical en reglas de reescritura, es decir en gramáticas independientes 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. Así que para este módulo creamos una gramática independiente del contexto para el español, una gramática computacional. Hacemos notar que existen diferencias entre las gramáticas teóricas y las computacionales. Por ejemplo, mientras escribir un elemento vacío en un árbol sintáctico no supone complicaciones para un lingüista teórico, para un lingüista computacional si hay complicaciones.

Este método de gramáticas independientes del contexto también lo empleamos en el método de obtención de los patrones de manejo, por lo que su construcción, que se describe en otras secciones de este capítulo, considera como meta ambos usos. Los detalles de compilación del diccionario y de la construcción del analizador para este modelo los presentamos en la sección 3.3 y secciones subsecuentes.

Proximidad semántica.

Este modelo está relacionado con el conocimiento semántico. Se requiere para desambiguar oraciones que son ambiguas, 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*.

No se trata de desambiguar el sentido mismo de las palabras. Esta tarea de desambiguación es distinta y se ha venido desarrollando como una subárea del procesamiento lingüístico de textos mediante computadora, considerando la desambiguación entre los sentidos dados en un diccionario, tesauro o similar. La desambiguación de sentidos de las palabras se ha estudiado con métodos estadísticos [Gale *et al*, 92], [Yarowsky, 92, 95], [Pedersen, 2000], métodos basados en conocimiento [Agirre & Rigau, 96], o con métodos mixtos [Jiang & Conrath, 97], [Rigau *et al*, 97]. Aunque se han alcanzado altos estándares, en esta desambiguación usualmente sólo se han seleccionado pequeños conjuntos de palabras con distinciones claras en el sentido.

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. Los detalles del uso de este modelo los presentamos en la sección 3.7.

Módulo de votación.

Para resolver la ambigüedad nos basamos en la asignación de pesos, o probabilidades de cada variante del análisis. En el caso ideal, una sola variante debería tener l como probabilidad y todas las demás variantes 0. En la práctica no podemos obtener una sola variante ya que ni aún los hablantes nativos pueden elegir una sola variante siempre.

En nuestra propuesta para desambiguación sintáctica de textos sin restricciones enfatizamos la necesidad de diversos modelos. Cada uno de los módulos de los modelos propuestos da como resultado una serie de variantes de análisis sintáctico de la oración de entrada. De entre todas las variantes resultantes nuestro modelo selecciona las más adecuadas.

En cada módulo de modelo la salida resultante es un grupo de distintas variantes que no están ordenadas. Para ordenarlas asignamos un peso a cada variante. Nosotros proponemos la asignación de pesos a las variantes de acuerdo a la complejidad y las características específicas de los métodos que las producen, así como una forma de compatibilidad. Sin una transformación a una forma compatible no sería posible determinar las variantes sobresalientes porque sus valores no serían comparables.

La asignación de pesos a las variantes dentro de cada modelo la realizamos de acuerdo a las características que la especifican y probabilidades a priori. Las características específicas corresponden al modelo mismo, qué tanto satisface la variante esas características. Las probabilidades corresponden al uso más frecuente de determinadas estructuras o determinadas realizaciones sintácticas. Estas posibilidades varían con cada modelo y la información disponible para ellos, pero en general, consideramos lo siguiente: enumeración de características distintivas del modelo, número de características o parámetros satisfechos dentro de cada modelo, diferenciación entre opciones en cada modelo y probabilidades de empleo de subestructuras.

En cada una de las secciones a continuación donde describimos cada uno de los módulos presentamos la descripción de las asignaciones de pesos para cada uno de los modelos. Ejemplos de las formas en las cuales se asignan los pesos y la votación, así como sus complejidades se detallan en la sección 3.8.

3.3 CREACIÓN DE LA GRAMÁTICA GENERATIVA EXPERIMENTAL

Las gramáticas independientes del contexto especifican cómo se forman las oraciones a partir de sus partes constituyentes y cómo se deriva la información asociada con cada oración (es decir, su interpretación) de la información de sus partes. En la creación de este tipo de gramática se considera la capacidad de tratar oraciones no conocidas previamente, es decir, de realizar una generalización con respecto a los datos considerados como base para desarrollar la gramática. Esta generalización hace que se prediga la *gramaticalidad*[33] de nuevas oraciones respecto al conjunto de reglas considerado.

La creación de este tipo de gramáticas implica tomar decisiones sobre dos requisitos que están en conflicto: la precisión y la cobertura. La precisión mide el grado de acierto de la gramática en lo que se refiere al análisis sintáctico. La cobertura gramatical mide la proporción de oraciones que reciben un tratamiento aceptable, generalmente respecto a un corpus de evaluación. Ambas propiedades son muy importantes, mientras más precisa es una gramática mejor es la calidad de sus análisis y mientras mayor cobertura tenga mayor la variedad de estructuras que trata la gramática.

El conflicto entre ambas propiedades se presenta cuando se quiere aumentar el rendimiento de ellas. Para mejorar la precisión hay que incorporar más restricciones a la gramática con lo que se tiende a perder cobertura, ya que las nuevas restricciones suelen rechazar algunas oraciones más o menos correctas que ya se aceptaban. [Pereira, 96] afirma que esto se debe a que las restricciones más poderosas son en realidad idealizaciones de la actuación (lo que se realiza) real de los hablantes, es decir, que la actuación es mucho más permisiva que la competencia (el conocimiento gramatical que se tiene).

[33] Que obedecen leyes gramaticales, sin conocimiento del mundo.

Por el otro lado, si se quiere mejorar la cobertura, se tiene que aumentar el número de reglas. Cuando una gramática alcanza un tamaño considerable, cada vez es más difícil de controlar y extender, ya que las nuevas reglas entran en interacciones complejas con las anteriores. Por lo que oraciones que antes no presentaban problemas producen varios análisis equivocados, es decir, aumenta la ambigüedad y decrece la precisión.

La gramática que necesitamos en este caso, dado que no es el método más importante, no requiere condiciones óptimas en cuanto a cobertura y precisión. Nuestra gramática pretende considerar las construcciones más comunes, que nos permita identificar el elemento rector en cada grupo y las relaciones sintácticas para el orden de palabras usual.

Para verificar la gramática, los elementos que más contribuyen son el marcaje de características morfológicas y la gramática misma, las cuales detallamos a continuación.

Marcas morfológicas

El marcaje de partes del habla o de categorías gramaticales (en inglés *POS tagging*) es útil para el análisis sintáctico. Conocer esta marca para una palabra específica ayuda a descartar la posibilidad de que esa misma palabra tenga otra categoría gramatical. La ambigüedad en categoría gramatical se refiere a que una palabra puede tener varias categorías sintácticas, por ejemplo *ante* puede ser una preposición o un sustantivo. La desambiguación de este marcaje es muy útil para reducir la cantidad de ambigüedad que tiene que enfrentar el analizador sintáctico.

El marcaje de partes del habla es la subárea del procesamiento lingüístico de textos por computadora que considera el estudio de métodos y algoritmos para reducir el porcentaje de ambigüedad de categorías. Los métodos que se han empleado se pueden clasificar en tres tipos: lingüísticos, estadísticos y aprendizaje mediante máquina. La mejor precisión en métodos lingüísticos corresponde a [Voutilainen, 94] con 99.3% aunque no todas las palabras están completamente desambiguadas, su defecto es la gran cantidad de tiempo que consume el desarrollar un buen modelo del lenguaje puesto que se requieren muchos años de recursos humanos. Los resultados producidos mediante métodos estadísticos han logrado entre 95% y 97% [Ludwig, 96] de palabras marcadas correctamente. Su defecto es la dificultad de estimar con precisión el modelo del lenguaje, puesto que es necesario estimar los parámetros del modelo como en las siguientes formas: la probabilidad de que cierta palabra aparezca con cierta marca o la probabilidad de que una marca sea seguida por otra marca específica.

Existen métodos híbridos que combinan diferentes aproximaciones, como ya habíamos mencionado, por ejemplo el uso de recursos basados en estadísticas y en conocimiento. En el tipo de aprendizaje mediante máquina los autores emplean

algoritmos de aprendizaje para adquirir el modelo del lenguaje a partir de un corpus de entrenamiento, por ejemplo, el algoritmo de [Brill, 95] es un aprendizaje manejado por los errores basado en transformaciones. Casi todos estos métodos mencionados se aplicaron exclusivamente al inglés, salvo [Padró, 97] que menciona la aplicación de su método al español aunque sin reportar la precisión exacta.

La desambiguación de las partes del habla implica al menos un análisis sintáctico parcial en muchos casos por lo cual no ha sido posible obtener una desambiguación total. En consecuencia una alternativa es marcar las categorías gramaticales en base a las características morfológicas de las palabras y dejar al análisis sintáctico la desambiguación correspondiente.

En esencia, el marcaje es el análisis morfológico. Sin este análisis, el análisis sintáctico es imposible. Pero al considerar todas las marcas morfológicas posibles de cada palabra, el análisis sintáctico usualmente da muchas variantes, ya que considera cada una de las marcas de cada palabra para empatarlas en las reglas de la gramática. Sólo una marca de todas las posibles de la palabra aparece en una variante.

La gramática que creamos se apoya en las marcas morfológicas que contienen las palabras del corpus[34] que consideramos. Este corpus no contiene desambiguación de POS por lo que el número de análisis es mayor. Esta aparente desventaja tiene su contraparte, si el desambiguador de POS no es de muy buena calidad, además de los resultados arriba indicados que muestran que no ha sido posible la desambiguación total, entonces ocasionará que no se realice el análisis sintáctico de algunas oraciones o que de antemano se orille a un análisis sintáctico incorrecto.

El corpus LEXESP tiene las categorías PAROLE [Civit & Castellón, 98]. La clasificación de categorías gramaticales en PAROLE la presentamos a continuación, donde se indican los rasgos considerados. Aunque, la posibilidad teórica de consideración de rasgos es mayor, aquí solamente consideramos las que se encuentran en el corpus.

1. Adjetivo (A).

Tipo		Grado	Género		Número		Caso	Función
Valor	Clave		Valor	Clave	Valor	Clave		
Calificativo	Q	0	Femenino	F	Singular	S	0	0
			Masculino	M	Plural	P		
			Común	C	Invariable	I		

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

Ejemplo: *frágiles* <AQ0CP00>

2. Adverbio (R)

Tipo		Tipo	Grado	Función
Valor	Clave			
General	G	0	0	0

Ejemplo: *no* <RG000>

3. Artículo (T)

Tipo		Género		Número		Caso
Valor	Clave	Valor	Clave	Valor	Clave	
Definido	D	Femenino	F	Singular	S	0
Indefinido	I	Masculino	M	Plural	P	0
		Común	C			

Ejemplo: *la* <TDFS0>

4. Determinante (D)

Tipo		Persona	Género		Número		Caso	Poseedor
Valor	Clave		Valor	Clave	Valor	Clave		
Demostrativo	D	1	Femenino	F	singular	S	0	0
Posesivo	P	2	Masculino	M	Plural	P		
Interrogativo	T	3	Común	C	Invariable	N		
Exclamativo	E							
Indefinido	I							

Ejemplo: *tal* <DD0CS00>

5. Sustantivo (N)

Tipo		Género		Número		Caso	Género semántico	Grado
Valor	Clave	Valor	Clave	Valor	Clave			
Común	C	Femenino	F	Singular	S	0	0	0
Propio	P	Masculino	M	Plural	P			
		Común	C	Invariable	I			

Ejemplo: *señora* <NCFS000>

6. Verbo (V)

Tipo		Modo		Tiempo	
Valor	Clave	Valor	Clave	Valor	Clave
Principal	M	Indicativo	I	Presente	P
Auxiliar	A	Subjuntivo	S	Imperfecto	I
		Imperativo	M	Futuro	F
		Condicional	C	Pretérito	S
		Infinitivo	N		
		Gerundio	G		
		Participio	P		

Persona	Número		Género	
	Valor	clave	Valor	Clave
1	Singular	S	Femenino	F
2	Plural	P	Masculino	M
3				

Ejemplo: *acabó* <VMIS3S0>

7. Pronombre (P)

Tipo		Persona	Género		Número	
Valor	Clave		Valor	Clave	Valor	Clave
Personal	P	1	Femenino	F	Singular	S
Demostrativo	D	2	Masculino	M	Plural	P
Posesivo	X	3	Común	C	Invariable	N
Indefinido	I					
Interrogativo	T					
Relativo	R					

Ejemplo: *ella* <PP3FS000>

8. Conjunciones (C)

Tipo		—	Posición
Valor	Clave		
Coordinada	C	0	0
Subordinada	S		

Ejemplo: *y* <CC00>

9. Numerales (M)

Tipo		Género		Número		Caso	Función
Valor	Clave	Valor	Clave	Valor	Clave		
Cardinal	C	Femenino	F	Singular	S	—	—
Ordinal	O	Masculino	M	Plural	P	0	0
		Común	C				

Ejemplo: *cinco* <MCCP00>

10. Preposiciones (SPS00). Ejemplo: *a* <SPS00>

11. Números (Z). Ejemplo: *5000* <Z>

12. Interjecciones (I). Ejemplo: *oh* <I>

13. Abreviaturas (Y). Ejemplo: *etc*. <Y>
14. Puntuación (F). Todos los signos de puntuación (.,:;-¡!'¿?"%). Ejemplo "." <Fp>
15. Residuales (X). Las palabras que no encajan en las categorías previas. Ejemplo: *sine* <X>

Un ejemplo de marcas en el corpus, es el siguiente, para la palabra *bajo* que puede ser tanto una forma verbal, como preposición, adverbio, sustantivo o adjetivo: bajar<VMIP1S0> bajo<SPS00> bajo<RG000> bajo<NCMS000> bajo<AQ0MS00>

El valor común de género se emplea tanto para femenino como para masculino, por ejemplo: *alegre*. El valor *invariable* en número se emplea tanto en singular como en plural, por ejemplo, el pronombre *se*.

Desarrollo y ampliación de cobertura de la gramática

La creación y cobertura de la gramática para sistemas computacionales no puede basarse en la literatura sobre lingüística teórica por la falta de explicitud, la falta de atención a detalles poco teóricos como nombres propios, fechas, etc., y porque además no se consideran los problemas de implementación en computadora (por ejemplo los movimientos de grupos de palabras en distintas posiciones en la oración).

Por un lado, el desarrollo de una gramática grande es extremadamente lento. No existen métodos para hacer eficiente la ingeniería de gramáticas [Erbach & Uszkoreit, 90]. Desde el punto de vista computacional sería deseable modulizar el desarrollo de la gramática [Volk, 92]. Sin embargo, las reglas son muy interdependientes, por ejemplo: los grupos verbales contienen grupos nominales, los grupos nominales pueden representarse mediante grupos verbales en infinitivo, etc. En la sección 3.4 presentamos los detalles de la compilación de la gramática.

Por otra parte, no hay un consenso general sobre la medición de la cobertura de una gramática. Los participantes del *Saarbrücken Grammar Engineering Workshop*[35] reportaron el tamaño de sus gramáticas en bytes, líneas de código, número de reglas, número de unificaciones, descripciones diferentes de nodos, y una lista de los fenómenos lingüísticos que cubrían. La GPSG [Gazdar *et al*, 85] ilustra que el número de reglas por sí mismo no es una buena medida, porque algunas reglas son equivalentes a un gran número de reglas de gramáticas independientes del contexto.

Por esta razón, para indicar la creación y cobertura de nuestra gramática, presentamos el conjunto de oraciones de prueba en el Apéndice A y a continuación

[35] *1st Workshop on Grammar Engineering: Problems and Prospects*, organizado en Junio de 1990 en Saarbrücken, Alemania, por Gregor Erbach and Hans Uszkoreit.

describimos las estructuras sintácticas que consideramos:

1. Estructuras de cláusulas. Entre ellas: cláusulas principales, cláusulas subordinadas, oraciones compuestas.
2. Frases.
3. Frases verbales, de verbos auxiliares y finitos.
4. Frases nominales. Consideramos frases simples, la modificación con frases preposicionales y con adjetivos, los infinitivos sustantivados, los sustantivos compuestos y los números.
5. Frases preposicionales. En distintas funciones: como objetos de verbos, como modificadores de sustantivos, adjetivos y adverbios, y como complementos.
6. Frases adjetivas, que modifican los sustantivos.
7. Frases adverbiales. Como modificador verbal, en todas las posiciones posibles. Como complemento.
8. Listas de cláusulas y de frases (nominales, preposicionales y adjetivas).
9. Otros fenómenos lingüísticos.
    - Concordancia. En el grupo nominal, entre sustantivos y adjetivos, y todas sus variantes. Entre grupo nominal como sujeto y verbo. Entre verbo auxiliar y los grupos: de participio, de sustantivo y de adjetivo.
    - Grupos de tiempo. Por ejemplo: hace un mes, una semana, todo el año.
    - Puntuación. Separando grupos y como enfatizadores.

También consideramos algunos fenómenos específicos como el caso: adjetivo *todo* - artículo - grupo nominal, por ejemplo *todos los niños de la calle*.

Una evaluación estadística basada en un corpus del español, la presentamos más adelante en la sección dedicada a la verificación de la gramática.

Mejora en la gramática

Las reglas que compilamos cubren las estructuras sintácticas antes descritas. La lista completa de las reglas la presentamos en la siguiente sección. En esta sección detallamos las mejoras que introducimos a nuestra gramática independiente del contexto del español. A continuación las enumeramos:

1. Reglas recursivas, por ejemplo para aceptar varios adjetivos consecutivos.
2. Convenciones para mejorar la capacidad expresiva y para una formulación más compacta, como la alternancia.

3. Convención de opcionalidad, para permitir varios constituyentes que a su vez no sean obligatorios.
4. Restricción de concordancia, la empleamos para evitar una clase de generación en exceso.
5. La inclusión del elemento rector, marcado con el signo "@".
6. La inclusión de relaciones sintácticas, por ejemplo un adverbio tiene una relación de modificación (mod) respecto a un verbo rector.
7. Inclusión de elementos de puntuación.
8. Inclusión de marcas semánticas. Marcamos grupos nominales con descripción semántica de tiempo. Por ejemplo: semana, año, etc.
9. Pesos estadísticos para graduar el número de reglas que se usan en el análisis.

Las tres primeras mejoras son muy comunes y simplifican la labor de la persona que elabora las reglas. La cuarta mejora es indispensable en un lenguaje con tanta flexión como el español. Las restantes mejoras no son muy comunes en este tipo de gramáticas.

Pocos estudios han considerado la inclusión del elemento rector, con la misma noción de las gramáticas de dependencias, por ejemplo [Collins, 99]. Algunos han considerado la inclusión del núcleo-*h*, por ejemplo [Pollard, 84], [Sikkel & Akker, 93], [Sikkel, 97]. Nuestra razón principal para incluir el elemento rector, en este contexto, es facilitar la conversión de un árbol sintáctico de constituyentes resultante de un análisis sintáctico mediante CFG a un árbol de dependencias correspondiente a la DG (*Dependency Grammar*) para la misma oración. Este procedimiento se detalla en la sección 3.5. Como ya mencionamos, la estructura de dependencias tiene la ventaja de mostrar las relaciones entre las palabras mismas de la oración.

Consideramos el marcado muy simple de grupos nominales de tiempo para detectar complementos circunstanciales. La idea general es poder identificar por anotación en el diccionario mediante marcas de descriptores semánticos, las subclases del tipo: tiempo, lugar, manera, etc. De esta forma es posible mejorar la precisión sin aumentar considerablemente el número de variantes generadas. Suponemos que un etiquetamiento mayor de lexemas en el diccionario, del tipo mencionado, hará más exitosa la desambiguación.

Siguiendo a [Jones, 94], que consideró puntuación como las marcas que no son léxicas y que se encuentran en los textos, los elementos de puntuación que incluimos son: coma, punto y coma, dos puntos, punto, signos de interrogación, signos de admiración, paréntesis, comillas, guiones, y apóstrofo. En nuestra gramática consideramos elementos de puntuación que funcionan como enfatizadores (*dijo que quería "un dulce"*), como separadores de listas de elementos similares (*rojo, verde,*

azul), y como delimitadores de modificadores (adverbios, circunstanciales). La inclusión de elementos de puntuación está relacionada a la calidad de la gramática y con la disminución de la cantidad de frases correctas que la gramática no puede analizar.

Partimos de una gramática general en base a manuales gramaticales y de un corpus de textos reales, pero tuvimos que reducir la generalidad de la gramática para evitar el elevado número de variantes. Por ejemplo, un complemento circunstancial puede estar realizado sintácticamente mediante adverbios, grupos preposicionales, grupos del gerundio; al considerar un complemento realizado como grupo nominal, se incrementa el número de variantes ya que cualquier grupo nominal sería considerado adicionalmente a su condición de posible sujeto, objeto directo o constituyente de grupos preposicionales y grupos de gerundio como un complemento circunstancial.

Es imposible atribuir valores absolutos: cierto o verdadero a la aplicabilidad de una regla y a la estructura gramatical resultante. No podemos partir de la suposición de que cada regla, aunque se haya mostrado su validez se pueda aplicar siempre en la misma forma. Por lo que es necesario considerar leyes probabilísticas. Así que asignamos pesos bajos (prioridad alta) a las reglas más empleadas y pesos más altos (prioridad baja) a las reglas que además de introducir un mayor número de variantes no son muy empleadas. Por ejemplo un grupo nominal algunas veces es un complemento circunstancial.

Verificación preliminar de la gramática

La verificación de una gramática, se realiza manualmente o semiautomáticamente mediante computadora. Lo cual es menos complicado cuando se trata de una gramática pequeña. Para verificar una gramática grande se ha considerado el empleo de un corpus, las objeciones son que un corpus no contiene de forma sistemática ejemplos de los fenómenos lingüísticos. [Gazdar, 99] considera los siguientes criterios como los adecuados para verificar una gramática computacional:

- Si la gramática genera en exceso, es decir, si la gramática acepta construcciones incorrectas.
- Si la gramática subgenera, es decir, si la gramática no puede analizar frases correctas.
- Si la gramática asigna estructuras apropiadas a las oraciones que logró analizar.
- Si la gramática es bastante simple.
- Si la gramática es general, es decir, que sea una gramática capaz de realizar generalizaciones con respecto a las estructuras consideradas.

Considerando estos criterios, creamos un corpus de oraciones de prueba. Los

conjuntos de pruebas se construyen desde el punto de vista lingüístico [Flickinger *et al*, 87], [Balkan, *et al*, 94], [Netter *et al*, 98]. También se ha intentado usar corpus [Balkan & Fouvry, 95], con distintos niveles de marcado, con el inconveniente de la cantidad de trabajo que se requiere para transformarlo, porque usualmente las oraciones de corpus no contienen fenómenos lingüísticos aislados ni variación sistemática de ellos. [Bröker, 2000] propone un método de reuso del conocimiento para construir la gramática con la finalidad de compilar su correspondiente conjunto de pruebas, el objetivo principal que logra es de cobertura de la gramática.

Nuestro conjunto de prueba, actualmente, cubre los fenómenos lingüísticos considerados en la gramática pero no tiene el propósito de cubrir toda la gramática. Su objetivo principal es mostrar lo siguiente:

- Ejemplos del tipo de construcciones que se analizan correctamente.
- Ejemplos negativos, es decir, para qué construcciones las oraciones se rechazan.
- Ejemplos de concordancia, ya que está explícita en las reglas, mostrar qué tipos de concordancia se consideraron.

Cada uno de los ejemplos tiene el propósito de mostrar un fenómeno lingüístico. Con el proceso de este corpus no se observan todos los resultados de las mejoras consideradas, ya que su función está enfocada a la calidad del analizador sintáctico: a la disminución de variantes, a la asignación de estructuras apropiadas, y a la simplificación de la gramática. Este corpus solamente es adecuado para indicar que la gramática está caracterizada para reconocer si las oraciones pertenecen al lenguaje descrito o no (adecuación observada).

A falta de una metodología aceptada de forma general para la medición del funcionamiento de una gramática, que sea objetiva, rigurosa y verificable, consideramos el uso de un conjunto de pruebas para mostrar la cobertura de la gramática y adicionalmente pruebas en un conjunto grande de textos. A continuación presentamos los resultados obtenidos con el analizador sintáctico que describimos en las secciones previas para el análisis sintáctico del corpus LEXESP.

En un fragmento del corpus, de 2Mby, se tienen 2552 oraciones. Del total de oraciones se analizó el 66%. La longitud promedio fue de 18 palabras, aunque el rango de longitud va de una palabra a 156 palabras. El número de variantes va de una variante a 10^9 variantes, con un promedio de 98×10^6 variantes. De las 872 oraciones que no se analizaron, 200 corresponden a frases donde faltaban marcas morfológicas.

De 119,007 oraciones, el 50% del corpus, se analizó el 55% de oraciones, con una longitud promedio de 22 palabras. El rango de longitud va de una palabra a 297 palabras. El número de variantes va de una variante a 10^{10} variantes, con un promedio de 129×10^6 variantes. Un total de 15,477 oraciones de las 54010 oraciones que no se analizaron son oraciones con palabras sin marca morfológica.

Como habíamos mencionado, crear una gramática computacional grande toma mucho tiempo. Entre las tareas que se deben realizar, está la modificación de las características de los datos de entrada. Entre los aspectos que consideramos con susceptibilidad de mejora se encuentran:

- Marcaje de POS de mejor calidad.

  Incluyendo el marcaje de POS de los casos acusativo y dativo.

  La generación en exceso de marcas de POS alimenta la generación excesiva de análisis sintácticos puesto que cada marca más de la necesaria genera al menos una marca sintáctica más de las necesarias.

- Inclusión de marcas semánticas. Por ejemplo locativo, etc.

- Modificación de las relaciones sintácticas. Por ejemplo: la reducción de relación de un adverbio, que ahora se considera tanto en relación adverbial como circunstancial.

Existen muchos otros factores que inciden en la verificación de la gramática que por el momento están fuera del ámbito de este estudio, como el género de los textos. Por ejemplo, oraciones muy largas de tipos específicos de texto son mucho más complicadas y difíciles de analizar, oraciones del lenguaje hablado, etc.

3.4 COMPENDIO DE REGLAS GRAMATICALES

En esta sección presentamos todas las reglas gramaticales que compilamos. En secciones anteriores dimos las razones teóricas, aquí presentamos los detalles y aspectos prácticos.

Para describir construcciones recursivas, empleamos reglas recursivas como la siguiente, donde el elemento de la izquierda también aparece en la parte derecha de la regla:

LIS_CLAUSE -> @:CONJ [SEP_O] coord_conj:LIS_CLAUSE

en esta regla, además, ejemplificamos la opcionalidad que se marca con corchetes. El elemento SEP_O (separadores en la oración) puede aparecer o no en una lista de oraciones precediendo a una conjunción. Con la regla anterior, describimos una lista de cláusulas que puede estar constituida por un sin fin de cláusulas, ya que adicionalmente contamos con la regla:

LIS_CLAUSE -> @:CLAUSE

que está embebida en la regla siguiente, donde se considera la lista de cláusulas separadas por elementos de puntuación:

LIS_CLAUSE -> @:CLAUSE [[SEP_O] coor_conj:LIS_CLAUSE]

Un ejemplo de las convenciones para mejorar la capacidad expresiva de la gramática y que permitirá tener una formulación más compacta, es la alternancia. En el siguiente ejemplo, el separador en las oraciones puede ser una coma, un punto y coma, dos puntos, etc.:

SEP_O → ',' | ':' | ';' | '...' | '(' | '-' | ')'

Las restricciones más comunes son las que tratan los fenómenos de concordancia y subcategorización. La subcategorización como ya vimos, es una

información que especifica las propiedades de combinación de las palabras. La subcategorización describe los requisitos sintácticos que impone un determinado elemento léxico sobre sus argumentos o complementos. La subcategorización solamente se considera en forma general, mediante reglas que indican la posibilidad de que un verbo tenga objeto directo, objeto indirecto u otros complementos. No se considera en detalle en esta gramática principalmente por su incapacidad para describirla relacionada a cada palabra y al contexto, además, como indicamos en el capítulo dos, esta información de una manera adecuada y mucho más completa se incluye en los patrones de manejo.

En lugar de unificación para los rasgos, explícitamente marcamos las características en las reglas. El poder de la unificación es la ventaja que ofrece en la creación o ingeniería de la gramática, al reducir la inmensa labor de especificarla, es decir, de marcarla. Según datos de [Uszkoreit & Zaenen, 96] la especificación de gramáticas grandes de unificación ha tomado alrededor de cuatro años, mientras que el desarrollo de gramáticas anotadas ha sido de ocho a doce años. Nosotros especificamos la concordancia en una forma general y un módulo de programación que desarrollamos la expande con las restricciones de rasgos. Por lo que aunque no es muy detallada, permitió que su elaboración se realizara en un tiempo mucho más corto.

La concordancia al igual que la subcategorización es difícil de expresar en gramáticas independientes del contexto debido a que estas características implican dependencia del contexto. Por ejemplo, hay que conocer las características del sujeto y del verbo principal para determinar si están en concordancia, es decir, se necesita consultar la información de varios constituyentes.

La solución que aplicamos para la concordancia es especificarla directamente en las reglas. Por ejemplo, la siguiente regla para el grupo de determinantes (DETER) abarca los determinantes (DET) como *esta*, *su*, etc. y los artículos (ART) como *el*, *un*, etc.:

DETER(nmb,gnd)

-> DET(nmb,gnd)

-> ART(nmb,gnd)

Estas dos reglas, las convertimos automáticamente en ocho reglas, ya que *nmb* representa el rasgo de número que tiene dos valores: singular y plural, y *gnd* que representa el rasgo género que tiene los valores: femenino y masculino. De la misma forma se modifican las reglas para incluir el número de persona en los grupos verbales y en los grupos nominales. La concordancia afecta únicamente el lado derecho de las reglas.

Aunque de esta forma tenemos un gran número de reglas, conseguimos disminuir la generación en exceso, además de que no es necesario escribir todas las

reglas por el proceso automático que las expande. Por ejemplo, en la siguiente regla el sustantivo (N) y el grupo del adjetivo (AP) deben concordar en género y número, en todas sus posibilidades. Para indicar que no se requiere la concordancia, se escriben variables diferentes: *nmb2*, *gnd1*, etc.

NOM(nmb,gnd,pers) -> @:N(nmb,gnd,pers) mod:AP(nmb,gnd)

En el compendio de reglas gramaticales independientes del contexto consideramos como punto principal especificar un conjunto de reglas que permitieran tener la mayor cobertura posible sin reconocer oraciones incorrectas del español. Como ya explicamos en la sección anterior, esto representa un conflicto. Nosotros optamos por asignar prioridades de tal manera que sólo se consideren las reglas que no son muy generales en caso de que no se pueda analizar con las reglas de mayor prioridad.

Estas reglas que consideramos no generales y que ocasionan errores en una mayor cantidad de oraciones, se refieren a casos como el de grupos nominales. Por ejemplo, *modelo Granada* (refiriéndose a un modelo de automóvil), *plan castración, pilas botón, etc.*, para analizarlos correctamente se requiere aumentar el conjunto de reglas con la siguiente regla donde no se unifican los rasgos:

(20) NP(nmb, gnd) → @:NP(nmb, gnd) NOM(nmb1, gnd1)

es decir, se eliminaría la concordancia, lo que equivaldría por un lado, a permitir concordancia incorrecta en los casos que por error se presentaran, o a cometer errores en la ocurrencia de elementos disjuntos (por ejemplo: *trasladaron a las pantallas fenómenos sociales, pusieron en el escenario rosas rojas*). Así que dimos una prioridad muy baja a esta regla. Las prioridades aparecen entre paréntesis al inicio de cada regla, por omisión la probabilidad es cero, que es la más alta prioridad.

En las siguientes secciones mostramos detalladamente la gramática independiente del contexto que compilamos para el español.

Signos convencionales de la gramática

Prioridades de las reglas

#---------- 0 – la mayor prioridad (construcciones sencillas)

#---------- 5 a 15 – construcciones complejas

#---------- 20 – la menor prioridad (grupo del sustantivo sin concordancia)

Categorías gramaticales

#---------- ADJ : adjetivo

#---------- ADV : adverbio

#---------- ADVP : grupo adverbial

#---------- AP : grupo del adjetivo

#---------- AUX : verbo auxiliar

#---------- BEG_S : puntuación inicio oración

#---------- CIR : complementos circunstanciales

#---------- CLAUSE : cláusula

#---------- CLAUSIN : cláusula sin circunstanciales

#---------- CONJ_C : conjunciones coordinantes

#---------- CONJ_SUB : conjunciones subordinantes

#---------- DETER : determinante

#---------- END_S : signos de puntuación al final oración

#---------- GER : gerundio

#---------- INFP : grupo del verbo en infinitivo

#---------- LIS_CLAUSE : lista de cláusulas

#---------- LIS_NP : lista de grupos nominales

#---------- LIS_PP : lista de grupos preposiciones

#---------- CONJ : conjunciones

#---------- N : sustantivo

#---------- N_TIE :sustantivo (descriptor semántico de *tiempo)*

#---------- NOM : grupo nominal sin determinante

#---------- NOM_TIE : grupo nominal sin determinante con descriptor semántico de *tiempo*

#---------- NP : grupo nominal

#---------- NP_TIE : gpo. nominal con descriptor semántico de *tiempo*

#---------- NUM : numeral

#---------- PART : participio

#---------- PP : frase preposicional

#---------- PPR :pronombre

#---------- PPR_C : pronombre acusativo y dativo

#---------- PPR_D : pronombre demostrativo

#---------- PPR_ID : pronombre indefinido

#---------- PPR_N : pronombre ordinal

#---------- PPR_IT : pronombre interrogativo

#---------- PPR_PE : pronombre personal

#---------- PPR_PO : pronombre posesivo

#---------- PPR_R : pronombre relativo

#---------- PR : preposición

#---------- S : oración completa de entrada

#---------- SEP_O : signo de puntuación dentro de la oración

#---------- VERB : verbo

#---------- VIN : verbo en indicativo

#---------- VCO : verbo en condicional

#---------- VSJ : verbo en subjuntivo

#---------- VP : grupo verbal finito

#---------- VP_DOBJ : objeto directo del verbo finito

#---------- VP_INF : grupo del verbo en infinitivo para no auxiliares

#---------- VP_INF_DOBJ : objeto directo del infinitivo

#---------- VP_INF_OBJS : secuencia de otros objetos del infinitivo

#---------- VP_MODS : modificador del verbo

#---------- VP_OBJS : secuencia otros objetos del verbo finito

#---------- VP_V : núcleo del grupo verbal

#---------- V_INF : núcleo del grupo del infinitivo

Títulos de relaciones y de parámetros

#---------- adver : relación adverbial

#---------- cir : relación circunstancial

#---------- comp : relación completiva

#---------- coord_conj : relación coordinada o conjuntiva

#---------- det : relación determinativa

#---------- dobj : relación objeto directo

#---------- gnd : parámetro de género

#---------- mean : parámetro de clasificación de verbo (*modal* o *aspectual*)

#---------- mod : relación modificadora

#---------- nmb : parámetro de número

#---------- pers : parámetro de persona

#---------- pred : relación predicativa

#---------- prep : relación prepositiva

#---------- subj : relación de sujeto

#---------- subor : relación de subordinación

Reglas de la gramática

S

-> [BEG_S] @:LIS_CLAUSE END_S # una o más CLAUSE

LIS_CLAUSE

-> [coor_conj:CONJ] @:CLAUSE [SEP_O coor_conj:LIS_CLAUSE]

 # ella dice, ella hace

-> coor_conj:LIS_CLAUSE [SEP_O] @:CONJ [SEP_O] coor_conj:LIS_CLAUSE

 # y ella busca ...

-> @:LIS_CLAUSE coor_conj:LIS_CLAUSE

#cuando llegaron el hecho estaba consumado

CLAUSE

-> [coor_conj:CONJ] @:CLAUSIN

-> @:CLAUSE [SEP_O] cir:CIR

 # El investigador descubre algunas cosas, de vez en cuando ....

-> cir:CIR [SEP_O] @:CLAUSE

 # Entre semana, por decisión del jefe, están restringidos

CLAUSIN

-> [subj:LIS_NP(nmb,gnd,pers)] @:VP(nmb,pers,mean)

 # El investigador descubre algunas cosas

-> [subj:LIS_NP(nmb,gnd,pers)] [SEP_O] [adver:ADVP [SEP_O]]

```
                @:VP(nmb,pers,gnd,AUX)
        #  Los sobres, comúnmente blancos, son ahora membretados
        -> subj:LIS_NP(nmb,gnd,pers) SEP_O [adver:ADV [SEP_O]]
           @:VP(nmb,pers,mean)
        # El investigador, frecuentemente descubre algunas cosas
SEP_O
        -> ',' | ':' | ';' | '...' | '(' | '"' | '-' | ')'
END_S
        -> '-'| ')'| '!'| '"'| '?'| '.'
BEG_S
        -> '-' | '¿' | '¡'
CONJ
        -> CONJ_C                                   # y, o, sino, pero
        -> CONJ_SUB                                 # si, porque sea, ya
(10)    -> @:CONJ_C CONJ_SUB                        # sino porque
        -> '...'
GERP
        -> @:GER                                    # caminando
        -> @:GER dobj:NP(nmb,gnd,pers) [obj:PP]     # brincando una barda
        -> @:GER obj:LIS_PP [dobj:NP(nmb,gnd,pers)] # caminando por el patio
LIS_GERP
        -> @:GERP [',' coord_conj:GERP]             # caminando, corriendo
        -> LIS_GER  @:CONJ coord_conj:GERP
        # caminando, corriendo y saltando
CIR
        -> @:ADVP    # mal, durante meses
        -> [mod:'todo' @:NP_TIE(nmb,gnd,pers)       # toda esta semana
        -> @:PR prep:NP_TIE(nmb,gnd,pers)           # y a los dos días
        -> @:PR prep:HACE_TIE                       # desde hace una semana
        -> @:LIS_GERP
        #  ... y maquinando trastadas en grupo
        -> @:LIS_PP [mod:ADV]
```

 # En Okinawa, a finales de la segunda guerra, cuando...

(20) -> @:LIS_NP(nmb,gnd,pers)

 # Dos edificios antes, junto a una tienda, venden ..

HACE_TIE

-> @:'hacer' NP_TIE(nmb,gnd,pers)

LIS_NP(nmb,gnd,pers)

-> @:NP(nmb,gnd,pers)

LIS_NP(PL,gnd,pers)

-> @:NP(nmb,gnd) ',' coord_conj:LIS_NP(nmb1,gnd1) # *bajo, gordo, rechoncho*

-> LIS_NP(nmb1,gnd1) @:CONJ coord_conj:NP(nmb,gnd)

 # la mezquindad, el afán crítico, y la envidia de sus semejantes

(10) -> LIS_NP(nmb1,gnd1,pers1) @:CONJ &coor_conj:PP

 # la mezquindad, el afán crítico, y hasta la envidia de sus semejantes

NP(nmb,gnd,pers)

-> [det:DETER(nmb,gnd)] @:NOM(nmb,gnd,pers)*#los científicos americanos*

-> @:PPR_ID(nmb,gnd,pers) [prep:PP] *# muchas | muchas de ellas*

-> @:PPR_IT(nmb,gnd,pers) [prep:PP] *# quién | quién de ellas*

-> @:PPR(nmb,gnd,pers) *# ella*

-> [det:DETER(nmb,gnd)] @:'cual' *# lo cual | las cuales*

-> [det:DETER(nmb,gnd)] @:PPR_PO(nmb,gnd,pers) *# lo suyo | las suyas*

-> [det:DETER(nmb,gnd)] @:PPR_N(nmb,gnd,pers) *# la primera*

-> mod:'todo' @:NP(nmb,gnd,pers) *# todos los mercados*

-> '“' @:NOM(nmb,gnd,pers) '”' *# "feliz"*

(10) -> &det:DETER(nmb,gnd) @:N(nmb,gnd,pers) pred:PP &comp:AP(nmb,gnd,pers)

 # un libro de cuentos desgastado por los años - aceite de oliva con residuos

-> [&det:DETER(nmb,gnd)] @:NOM(nmb,gnd,pers) [','] pred:LIS_PP [',']

 # el primer día de sol y de viento

-> @:DETER(nmb,gnd) pred:PP *# el de las rosas*

-> mod:AP(nmb,gnd) @:NOM(nmb,gnd,pers) *#amplias zonas de árboles*

(5) -> det:DETER(nmb,gnd) @:AP(nmb,gnd) [pred:PP] *# el rojo*

(20) -> @:NOM(nmb,gnd,pers) mod:NOM(nmb1,gnd1,pers1) # *pilas botón*

NOM(nmb,gnd,pers)

-> [num:NUM(nmb)] @:N(nmb,gnd,pers) # *5000 años*

-> @:N(nmb,gnd,pers) [','] mod:AP(nmb,gnd) [','] # *noticiario, televisivo,*

-> mod:AP(nmb,gnd) @:N(nmb,gnd,pers) # *alguna galaxia*

-> @:N(nmb,gnd,pers) pred:PP # *aceite de oliva*

(15) -> @:N(nmb,gnd,pers) comp:N(nmb,gnd,pers) [mod:AP(nmb,gnd)]

 # tiempos más lejanos

-> mod:AP(nmb,gnd) @:N(nmb,gnd,pers) mod:AP(nmb,gnd) # *única mano válida*

-> NUM(nmb) # *5000*

-> INFP # *comprar una torta, beber un jarrito y escuchar rock*

PPR(nmb,gnd,pers)

-> PPR_D(nmb,gnd,pers) # *éste | éstos*

-> PPR_PE(nmb,gnd,pers) # *ello | él*

-> PPR_R(nmb,gnd,pers) # *cuya | mismo*

DETER(nmb,gnd)

-> DET(nmb,gnd) # *aquel*

-> ART(nmb,gnd) # *el, un*

AP(nmb,gnd)

-> @:ADJ(nmb,gnd) comp:ADJ(nmb,gnd) # *antitelevisiva tradicional*

-> @:ADJ(nmb,gnd) adver:ADV # *racial extremadamente*

-> mod:ADV @:ADJ(nmb,gnd) # *muy feliz*

-> @:ADJ(nmb,gnd) [',' comp.:AP(nmb,gnd)] # *racial, sexual o física*

-> AP(nmb,gnd) @:CONJ coor_conj:ADJ(nmb,gnd) # *vertical u horizontal*

-> @:ADJ(nmb,gnd) pred:LIS_PP # *lleno de ...*

PP

-> @:PR prep:LIS_NP(nmb,gnd,pers) # *de la tal señora*

> @:QUE

> @:PR prep:QUE # *de que se enojaba*

> @:PR prep:INFP # *de caminar una hora*

(10) -> @:PR prep:CLAUSE # *de no se que señora*

QUE

-> @:'que' prep:CLAUSE # *que se enojaba*

-> @:'que' prep:NP(nmb,gnd,pers) # *que la señora*

LIS_PP

-> @:PP [',' coord_conj:LIS_PP] # *en noticiarios televisivos, en diarios, en ..*

-> LIS_PP @:CONJ coord_conj:PP # *al patio trasero y a la escalera*

(30) -> @:CONJ coord_conj:PP ',' # *ni en espectáculos, ni en conseguir que.....*

ADVP

-> ADV # *bueno | malo*

-> @:PR adver:ADV [mod:ADV] # *por atrás*

-> @:ADV comp:NP_TIE(nmb,gnd,pers1) # *durante meses*

-> @:ADV mod:ADJ(nmb,gnd) # *tanto mejor*

-> @:HACE_TIE # *hace un año*

-> @:ADV adver:ADV # *incluso más*

(10) -> @:ADV comp:NP(nmb,gnd,pers) comp:QUE_NP # *más bajo que alto*

-> @:ADV mod:PP # *incluso de día*

(10) -> @:PP # *a decir verdad*

(10) -> @:ADV comp:NP(nmb,gnd,pers) # *como un rosario*

(20) -> @:ADJ # *feliz*

QUE_NP

-> @:'que' prep:NP(nmb,gnd,pers) # *que aquel hombre*

NP_TIE(nmb,gnd,pers)

-> [[mod:'todo'] det:DETER(nmb,gnd)] @:NOM_TIE(nmb,gnd,pers) # *todo el día*

-> det:DETER(nmb,gnd) @:NOM_TIE(nmb,gnd,pers) prep:PP # *el día de la bandera*

NOM_TIE(nmb,gnd,pers)

-> cuant:NUM(nmb) [mod:AP(nmb,gnd)] @:N_TIE(nmb,gnd,pers) [mod:AP(nmb,gnd)]

 # dos largos años grises

N_TIE(nmb,FEM,3PRS)

-> @:'semana'| @:'hora'| @:'mañana'| @:'tarde'| @:'noche'

N_TIE(nmb,MASC,3PRS)

-> @:'día'| @:'año'| @:'mes'| @:'ayer'| @:'siglo'| @:'minuto'| @:'milenio'| @:'decenio

-> @:'lunes'| @:'martes'| @:'miércoles'| @:'jueves'| @:'sábado'| @:'domingo'

-> @:'febrero'| @:'enero'| @:'marzo'| @:'abril'| @:'mayo'| @:'junio'

-> @:'julio'| @:'agosto'| @:'septiembre'| @:'octubre'| @:'noviembre'| @:'diciembre'

#**********

Grupo del verbo

#**********

VP_MODS

-> ADVP

-> @:LIS_GERP

VP(nmb,pers,gnd,AUX)

-> [clit:PPR_C(nmb1,gnd1,pers1)] @:VERB(nmb,pers,AUX) [mod:ADV] [dobj_suj:NP(nmb,gnd,pers)]

 # era pariente de ...

-> [clit:PPR_C(nmb1,gnd1,pers1)] @:VERB(nmb,pers,AUX) [mod:ADV] dobj:AP(nmb,gnd)

 # es fatal

-> @:VERB(nmb,pers,AUX) [mod:ADV] dobj:N(nmb,gnd,pers) obj:PP

 # hay vida en alguna...

-> @:VERB(nmb,pers,AUX) [mod:ADV] obj:PP dobj:N(nmb,gnd,pers)

 # hay en algún lugar una escuela...

VERB(nmb,pers,AUX)

-> VIN(nmb,pers,AUX)| VCO(nmb,pers,AUX)| VSJ(nmb,pers,AUX)

-> [clit:PPR_C(nmb1,gnd1,pers1)] @:'haber' [adver:ADVP] PART(SG,MASC,AUX) PART(nmb,gnd)

 # le había sido visto

-> @:'haber' aux:NP(nmb,gnd,3prs) *# había testigos*

VP(nmb,pers,mean)

-> VP_DOBJ(nmb,pers,mean)

-> VP_OBJS(nmb,pers,mean)

VP_DOBJ(nmb,pers,mean)

-> @:VP_OBJS(nmb,pers,mean) obj:LIS_NP(nmb1,gnd1,pers1)

 # clavaban sus dardos

-> @:VP_DOBJ(nmb,pers,mean) comp:LIS_PP

 # trasladó su fábrica a la frontera

-> @:VP_DOBJ(nmb,pers,mean) mod:VP_MODS

 # ordenó una fila moviendo las sillas

SUJ_DOBJ

->@:'al' prep:NP(nmb,gnd,pers)

->@:'a' prep:NP(nmb,gnd,pers)

->@:NP(nmb,gnd,pers)

VP_OBJS(nmb,pers,mean)

-> [adver:ADV] @:VP_V(nmb,pers,mean) [mod:VP_MODS] *#provocaban en su mente*

-> [adver:ADV] @:VP_V(nmb,pers,mean) [obj:LIS_PP] *#salieron del corral*

-> @:VP_OBJS(nmb,pers,mean) obj:LIS_PP

 # clavaban sus dardos por todo el cuerpo

-> @:VP_OBJS(nmb,pers,mean) mod:VP_MODS

 # jugaban el último partido provocándose a cada momento

VP_V(nmb,pers,mean)

-> [clit:PPR_C(nmb1,gnd1,pers1) [clit:PPR_C(nmb2,gnd,pers2)]] @:VP_SV(nmb,pers,mean)

 # se les llamase, se les haya dicho

VP_SV(nmb,pers,mean)

-> @:VERB(nmb,pers,mean) *# creo*

-> @:'haber'(nmb,pers) [adver:ADVP] PART(SG,MASC) *# habían incluso dudado*

-> @:'estar' [&adver:ADVP] AP(nmb,gnd) *# estaba contento*

-> @:'estar'(nmb,pers) [adver:ADVP] PART(nmb,gnd) *# está mal visto*

-> @:'ser'(nmb,pers) [adver:ADVP] PART(nmb,gnd)

 # es folicularmente handicapado

PPR_PE(nmb,gnd,3PRS)

(10) -> 'usted'

VERB(nmb,pers,mean)

-> VIN(nmb,pers,mean)| VCO(nmb,pers,mean)| VSJ(nmb,pers,mean)

PPR_PE(nmb,gnd,3PRS)

-> 'usted'

#**********

INFP

-> @:VP_INF [SEP_O coord_conj:INFP] *# cantar, reír*

-> INFP @:CONJ coord_conj:VP_INF *# vivir y morir*

-> [adver:ADV] @:V(INF,AUX) [adver:ADV] *# morir también*

VP_INF

-> @:VP_INF_DOBJ *# convertir la bandera de los rayos en oficial*

-> @:VP_INF_OBJS *# ir a la cárcel...*

VP_INF_DOBJ

-> @:VP_INF_OBJS [','] dobj_suj:SUJ_DOBJ [dobj_suj:SUJ_DOBJ]

 # dar su consentimiento

-> @:VP_INF_DOBJ [','] obj:LIS_PP

 # introducir unos centímetros en su interior

-> @:VP_INF_DOBJ [','] mod:VP_MODS

 # decir una palabras negando su sentir

VP_INF_OBJS

-> @:V_INF *#esperar pacientemente*

-> @:VP_INF_OBJS [','] obj:LIS_PP *# marchar hasta ...*

-> @:VP_INF_OBJS [','] mod:VP_MODS *#marchar torciendo ...*

V_INF

-> [adver:ADV] @:V(INF,mean) [adver:ADV] *# no estar hoy*

-> [adver:ADV] @:'haber'(INF) [adver:ADV] PART(SG,MASC) [adver:ADV]

 # no haber presentado puntualmente, habían siempre quedado..

-> [adver:ADV] @:'ser'(INF) [adver:ADVP] PART(nmb,gnd) [adver:ADV]

ser entrevistada

3.5 ALGORITMO DE TRANSFORMACIÓN DE ÁRBOLES DE CONSTITUYENTES A ÁRBOLES DE DEPENDENCIAS

Condiciones de transformación

La más importante de las mejoras introducidas en las gramáticas independientes del contexto (GPSG, HPSG) es la marca del elemento rector. Esta marca permite transformar, mediante un algoritmo, el árbol de constituyentes a un árbol simple de dependencias. Esta transformación permitirá simplificar la labor de identificación de valencias tanto en el analizador sintáctico general como para la compilación de los patrones de manejo (que se describe en el siguiente capítulo). Por ejemplo, en la siguiente regla para un grupo nominal, el sustantivo es el elemento rector.

NOM(nmb,gnd,pers) -> @:N(nmb,gnd,pers) Adj(nmb,gnd)

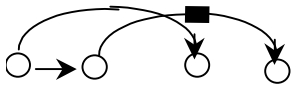
Dos conceptos son importantes en esta transformación.

1. La primera consideración inevitable para esta transformación, es la suposición de que todas las oraciones sujetas al análisis son proyectivas. [Mel'cuk, 88] indica que existen oraciones que no son proyectivas pero que todas ellas de alguna forma están marcadas de forma enfática, estilística, comunicativa o contienen elementos sintácticos especiales.

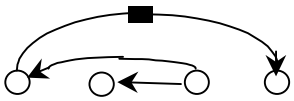
La proyectividad[36] es una propiedad del orden de palabras. Una oración se dice proyectiva si y solo si entre los arcos de dependencia que enlazan sus formas de palabras:

[36] También definida como adyacencia por algunos autores.

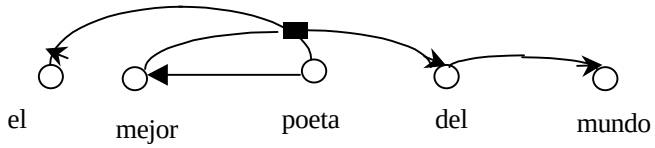
a) Ningún arco atraviesa a otro arco.



b) Ningún arco cubre el nodo tope.

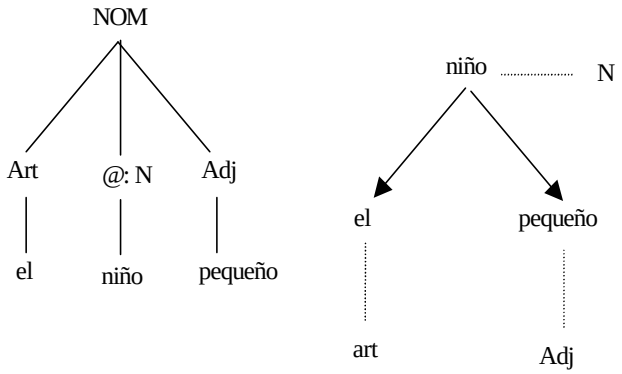


Un ejemplo de frase que viola la proyectividad es: *el mejor poeta del mundo.*



En las gramáticas independientes del contexto la proyectividad es estricta, es decir, es propiedad inalienable. La misma restricción se presenta también en algunos sistemas de análisis sintáctico basados en dependencias [Sleator & Temperley, 93], [Moortgat, 94], [Eisner, 96]. Así que nuestra consideración no es arbitraria.

2. La segunda consideración inevitable es que considerando la regla anterior, para transformar su árbol de constituyentes, tomamos el elemento rector como nodo raíz, y todos los nodos hijos restantes del elemento izquierdo de la regla como dependientes directas de él. Cada constituyente con *n* hijos contribuye con *n-1* dependientes.



Esto significa que podemos usar reglas con un solo núcleo, por ejemplo del tipo:

PP → @:PR N y CLAUSE → @:V NP

Donde PP es una frase preposicional, PR es una preposición, CLAUSE es una oración, V es un verbo y NP es un grupo del sustantivo. Los núcleos se marcan con el símbolo @. Estas reglas cumplen con la Forma Normal de Chomsky, que se detalla en la sección 3.6. En cambio, no es posible usar reglas del tipo:

CLAUSE → @:V @:PR N

donde N depende de PR y todo este grupo depende de V.

Estas dos consideraciones son necesarias y suficientes para hacer la transformación de árboles de dependencias a árboles de constituyentes. Ambos formalismos, dependencias y constituyentes describen el mismo lenguaje aunque las transformaciones no son de uno a uno. Pero el hecho de que un árbol de dependencias represente a varios árboles de constituyentes no viola la consideración de que se trata del mismo lenguaje. Después de la transformación se pueden identificar las estructuras iguales. Considerando restricciones no muy estrictas, cada gramática de constituyentes tiene una gramática de dependencias propia. Con esta condición podemos tomar para estudios teóricos y prácticos cualquier gramática.

La indicación de sentido de las flechas, es decir, la marca de dependencia, se define con las etiquetas que establecen las relaciones. Estas etiquetas son de modificación (*mod*), prepositivas (*prep*), etc.

Algoritmo básico de transformación

En el algoritmo de transformación de un árbol de constituyentes a uno de dependencias, se recorre el árbol de constituyentes en un orden determinado,

mediante un recorrido en profundidad. Empieza en la raíz y visita recursivamente a los hijos de cada nodo en orden de izquierda a derecha. Una vez que llega a nodos cuyos hijos cubren terminales, por ejemplo N(PL,FEM,3PRS) -> *NCFP000, asigna un nodo del árbol de dependencias al terminal del elemento rector y enlaza a los hermanos en el árbol de constituyentes como dependientes del nodo previamente definido en el árbol de dependencias.

Para cada uno de los nodos dependientes, se traslada su marca de dependencia para indicar la flecha de esa dependencia. El nodo superior del nodo de constituyentes se asigna como nodo superior del nodo rector definido en el árbol de dependencias y se elimina el nodo de constituyentes. De esta forma se va convirtiendo el árbol de constituyentes a uno de dependencias en forma ascendente. El último paso corresponde al enlace del nodo rector en el tope del árbol de constituyentes, ya que convierte al nodo raíz en un nodo que cubre terminal por lo que se detiene el proceso. En la **Figura 18** presentamos el algoritmo recursivo desarrollado.

Como ejemplo de esta transformación presentamos la transformación de una frase del corpus LEXESP. En la Figura 19 presentamos la representación que del árbol de constituyentes obtenemos con nuestra gramática generativa para la frase *Los alumnos solicitaron becas al director*. Con sangrías en el texto se marcan las agrupaciones. Los números de la izquierda corresponden a un número de orden

Convertir_a_dependencias
Para cada hijo **q**, del nodo **n** (del árbol de constituyentes) que no cubre un terminal,
de izquierda a derecha, **hacer lo siguiente:**
Convertir_a_dependencias
Asignar el nodo **m** (del árbol de dependencias) al elemento rector de los hijos del nodo **n**
Para todos los hijos del nodo **n** (que no sean el elemento rector) hacerlos dependientes de **m**
Trasladar las marcas de dependencias
Asignar como nodo superior de **m** al mismo nodo superior de **n** y eliminar el nodo **n**

Figura 18 Algoritmo de transformación de un árbol de constituyentes a uno de dependencias

alfabético de las reglas de la gramática. Las reglas que en la parte derecha sólo tienen un terminal entre paréntesis y asterisco inicial, como PR → <*SPCMS>, indican al

final la palabra que representan, tanto la cadena de entrada como la forma base, por ejemplo: (*solicitaron: solicitar*).

El árbol de constituyentes es la variante número 2 que obtuvimos con la siguiente entrada del corpus LEXESP:

Los alumnos solicitaron becas al director.

```
2:

7120      S -> @:CLAUSE $PERIOD

12576       CLAUSE ->  (subj) NP(PL,MASC,3PRS) @:VP_DOBJ(PL,3PRS,MEAN)

6924          NP(PL,MASC,3PRS) ->  (det) ART(PL,MASC) @:N(PL,MASC,3PRS)

307           ART(PL,MASC) -> <*TDMP0> ( Los: el, 0/0)

174           N(PL,MASC,3PRS) -> <*NCMP000> ( alumnos: alumno, 1/0)

26765         VP_DOBJ(PL,3PRS,MEAN) -> @:VP_DOBJ(PL,3PRS,MEAN)  (mod)
PP

23435         VP_DOBJ(PL,3PRS,MEAN)  ->  @:VIN(PL,3PRS,MEAN)  (obj)
N(PL,FEM,3PRS)

398             VIN(PL,3PRS,MEAN) -> <*VMIS3P0> ( solicitaron: solicitar, 2/0)

170             N(PL,FEM,3PRS) -> <*NCFP000> ( becas: beca, 3/0)

16763             PP -> @:PR  (prep) N(SG,MASC,3PRS)

301             PR -> <*SPCMS> ( al: al, 4/1)

175             N(SG,MASC,3PRS) -> <*NCMS000> ( director: director, 5/0)

142     $PERIOD -> <*Fp> ( .: ., 6/0)
```

Figura 19 Análisis sintáctico de constituyentes para la frase: *Los alumnos solicitaron becas al director*.

Los (0) el (*TDMP0) (0) él (*PP3MP000) (1) lo (*NCMP000) (2) los (*NP0000) (3)

alumnos (1) alumno (*NCMP000) (0)

solicitaron (2) solicitar (*VMIS3P0) (0)

becas (3) beca (*NCFP000) (0)

al (4) al (*NP00000) (0) al (*SPCMS) (1)

director (5) director (*NCMS000) (0) director (*NP00000) (1)

. (6) . (*FP) (0)

dónde entre paréntesis se presentan las marcas morfológicas codificadas de acuerdo al código *parole* empleado para el proyecto EURO WORDNET [Rodríguez *et al*, 98]. El primer número entre paréntesis corresponde al número de palabra en la oración de entrada, y los siguientes paréntesis con números corresponden a la numeración de diferentes marcas morfológicas. El asterisco marca su condición de terminal, es decir, de palabra de entrada.

En la Figura 20 presentamos la estructura de dependencias en la representación

```
2:
398  VIN(PL,3PRS,MEAN) -> <*VMIS3P0> ( solicitaron: solicitar, 2/0)
170     N(PL,FEM,3PRS) ->  (obj) <*NCFP000> ( becas: beca, 3/0)
301     PR ->  (mod) <*SPCMS> ( al: al, 4/1)
175        N(SG,MASC,3PRS) ->  (prep) <*NCMS000> ( director: director, 5/0)
174     N(PL,MASC,3PRS) ->  (subj) <*NCMP000> ( alumnos: alumno, 1/0)
307        ART(PL,MASC) ->  (det) <*TDMP0> ( Los: el, 0/0)
142     $PERIOD -> <*Fp> ( .: ., 6/0)
```

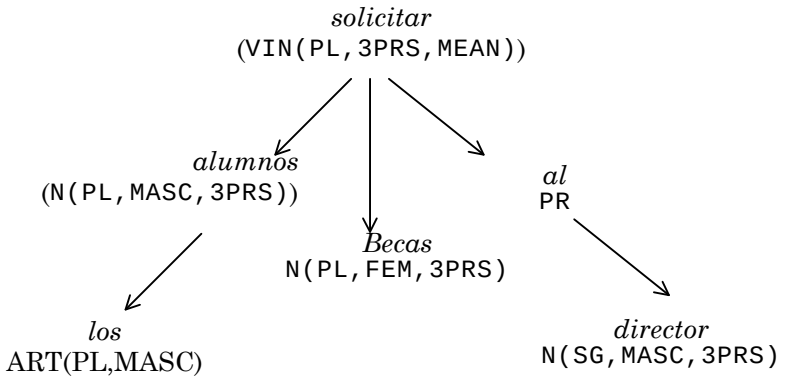


Figura 20 Análisis sintáctico de dependencias para la frase *Los alumnos solicitaron becas al director*.

obtenida con nuestra gramática generativa y enseguida en forma gráfica, esta estructura la obtenemos con el algoritmo desarrollado, a partir de la estructura de constituyentes de la Figura 19.

3.6 CONSIDERACIÓN DE LAS REGLAS PONDERADAS

El algoritmo empleado para realizar el análisis sintáctico con las reglas ponderadas relaciona cadenas de símbolos con el conocimiento lingüístico almacenado en las reglas y el diccionario de palabras marcadas. Este algoritmo es el mecanismo computacional que infiere la estructura de las cadenas de palabras a partir del conocimiento almacenado.

Un algoritmo de análisis sintáctico de este tipo es un procedimiento que prueba diferentes formas de combinar reglas gramaticales para encontrar una combinación que genere un árbol que represente la estructura de la oración de entrada para su interpretación correcta. Durante el procesamiento de los datos se crean muchas estructuras temporales, las estructuras finales son el resultado del análisis. Los algoritmos de análisis sintáctico más empleados por su eficiencia, se basan precisamente en gramáticas independientes del contexto.

Los algoritmos deciden qué reglas probar y en qué orden, para lo cual combinan diferentes estrategias y estructuras temporales. Existen diferentes estrategias: dirigido por las hipótesis o por los datos, procesamiento secuencial o paralelo, análisis determinista o no determinista. Las estructuras están relacionadas directamente con las estrategias empleadas.

El análisis sintáctico dirigido por las hipótesis o por la gramática es conocido también como descendente. Busca primero en la gramática las reglas y va construyendo estructuras hasta completar las palabras de la secuencia de entrada. Va construyendo estructuras desde el símbolo inicial ***S*** correspondiente a la oración, hacia abajo, hasta encontrar la secuencia de palabras de la entrada. El análisis sintáctico dirigido por los datos es conocido como ascendente. Parte de las palabras de la secuencia de entrada para ir encontrando las reglas cuya parte derecha contengan esas combinaciones de palabras adyacentes. Va construyendo estructuras hacia arriba hasta llegar al símbolo inicial que representa a la oración.

El procesamiento secuencial prueba una opción hasta el final y si falla regresa a puntos anteriores del proceso e incluso hasta el punto inicial. El procesamiento paralelo prueba diferentes posibilidades al mismo tiempo. Mientras el primero opera en una sola computadora, el segundo requiere procesamiento paralelo. Existen procesos que podrían considerarse intermedios entre éstos. Por ejemplo el pseudo paralelismo [Tomita, 86], que a partir de determinados puntos del proceso prueba diferentes opciones en secuencia y hasta resolver el conflicto continúa. Los algoritmos para el procesamiento paralelo son más complejos y difíciles de escribir, además de que requieren grandes cantidades de tiempos de cálculo, por lo que se han empleado escasamente.

El análisis determinístico sigue siempre un solo camino, mientras que el no determinístico tiene que elegir, en algunos puntos, diferentes caminos. Los algoritmos determinísticos son más eficientes aunque más limitados ya que no hay opciones. El que los algoritmos sean determinísticos o no determinísticos depende de la gramática y del analizador.

Los métodos ascendentes aprovechan su conocimiento de los elementos léxicos mientras que los descendentes aprovechan su conocimiento de las reglas gramaticales. Aunque los descendentes tienen la ventaja de considerar el contexto izquierdo, tienen la desventaja de considerar palabras y categorías que no aparecen en la secuencia de entrada y de repetir análisis con el mismo símbolo si éste aparece en distintos contextos. Los ascendentes tienen la ventaja de solamente considerar las palabras que aparecen en la cadena de entrada y de construir un análisis parcial de la misma estructura, pero tienen la desventaja de no tener restricciones contextuales por lo que prueban muchas combinaciones para las que no hay reglas.

Además de la estrategia, que tiene sus ventajas y desventajas, para tener un análisis eficiente un punto muy importante es almacenar los resultados intermedios para evitar la redundancia en el espacio de búsqueda. [Kay, 73, 80] propuso una estructura de datos que remediara esta falla, a la que denominó *chart* y que nosotros nombraremos *tabla* simplemente. Un *chart* o tabla es una estructura de datos que almacena resultados parciales, de manera que el trabajo no tenga que duplicarse. Lleva el registro de todos los constituyentes derivados en un punto del análisis, así como aquellas reglas que se han aplicado con éxito parcial pero que todavía no se logra completar. Estas estructuras generalmente se han representado mediante arcos. Los arcos activos son aquellos a los que les falta algún constituyente por reconocer. En la **Figura 21** mostramos la correspondencia entre las representaciones de árbol y de tabla para un grupo nominal.

La operación básica de un analizador sintáctico basado en tabla es combinar un constituyente incompleto con un constituyente ya completo. Por ejemplo un grupo nominal sin modificador posterior con un grupo preposicional complemento de sustantivo. El resultado será entonces un nuevo constituyente (un grupo nominal para este ejemplo) o un nuevo arco activo que es una extensión del anterior (un grupo

nominal al que le sigue faltando un modificador posterior). Todos los constituyentes que están completos se guardan en una lista hasta que son requeridos por el analizador

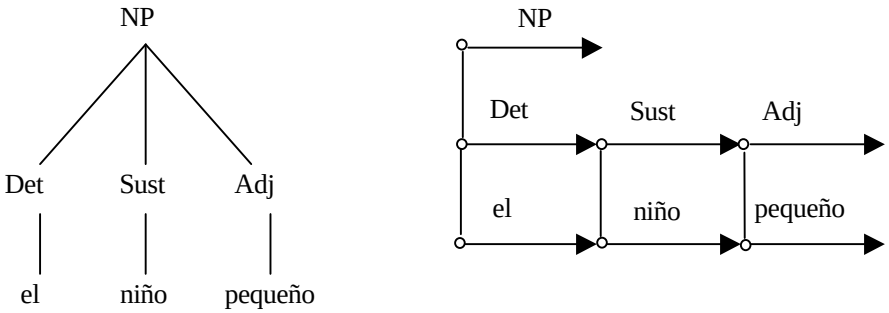


Figura 21. Representaciones de árbol y de tabla para el grupo nominal
El niño pequeño.

de tabla. Cuando el último arco activo se completa, termina el reconocimiento.

[Allen, 95] considera que los analizadores sintácticos basados en tabla son más eficaces que los que se basan solamente en búsqueda ascendente o descendente debido a que un mismo constituyente nunca se construye más de una vez, aunque claro está que la eficiencia práctica dependerá de la forma en que se implemente, las estructuras de datos que se empleen, el lenguaje de programación y la máquina específica.

Dadas las restricciones que impone de por sí una gramática independiente del contexto, escogimos este algoritmo de tabla como un analizador eficiente y simple para nuestro objetivo. Uno de los algoritmos ascendentes mejor conocidos por su poder para analizar cualquier gramática independiente del contexto es el algoritmo CKY [Kasami, 65; Younger, 67] pero no es eficiente. En cambio con manejo de tabla es muy conocido por su eficiencia e implementación [Eisner, 96], [Sikkel & Akker, 93]. Para los teóricos la eficiencia se refiere a que en el peor caso requiere un tiempo $O(n^3)$ y para una gramática de tamaño fijo requiere un tiempo $O(n^2)$ [Kay, 80], donde n es el número de palabras de la oración.

El algoritmo CKY emplea una gramática en forma especial, la Forma Normal de Chomsky (CNF en inglés). En las gramáticas CNF, las reglas de producción son del tipo $A \rightarrow BC$ o $A \rightarrow a$. Cualquier gramática independiente del contexto en CNF puede generar un lenguaje independiente del contexto. Para convertir una gramática independiente del contexto a la forma normal de Chomsky se requieren los siguientes pasos:

1. Añadir un nuevo símbolo inicial.

2. Eliminar todas las reglas con el elemento vacío (ε).
3. Eliminar todas las reglas de un solo elemento en la derecha (A → B, A → A).
4. Convertir todas las reglas restantes.
   - Introducir auxiliares por terminales (en lugar de A → d B, introducir A → Z B y Z → d).
   - Introducir auxiliares por no-terminales (en lugar de A → BCD, introducir A → B X y X → CD).

El algoritmo CKY opera de la siguiente forma. Se considera una gramática CNF con *k* no-terminales, *m* terminales y *n* reglas de producción. Para saber si una cadena de entrada puede ser generada por esa gramática, hace lo siguiente: si *a*[i, j] es una subcadena de la entrada desde la posición *i* hasta la *j* (0< *i*, *j* <= *n*), construye una tabla que diga para cada *i* y *j*, cuál símbolo (si existe) genera la cadena *a*[i, j]. Una vez que tiene la tabla, revisa si el símbolo inicial puede generar la cadena de entrada *a*[1, n].

La tabla se construye por inducción en la longitud de las subcadenas *a*[i, j]. Es fácil para subcadenas de longitud 1: A genera *a*[i, i+1] = *a*, si y sólo si existe la regla A → *a* en la gramática. Para longitudes mayores se hace una revisión exhaustiva para cada regla de producción. Para la regla A → BC revisa si existe una *k* (entre *i* y *j*) tal que B genera *a*[i, k] y C genera *a*[k+1, j]. Como estas subcadenas son menores que *a*[i, j] se encuentran ya en la tabla.

El algoritmo CKY se muestra en la Figura 22 [Goodman, 98]. La estructura de datos de la tabla es un arreglo booleano de tres dimensiones, donde un elemento chart[i, A, j] es verdadero si existe la derivación $A \overset{*}{\Rightarrow} \omega_i, \ldots \omega_{j-1}$, de lo contrario es falso. En la Figura 22 la línea

chart [s, A, s+1] := chart [s, A, s+ l] ∨ chart [s, B, s+ t] ∧ chart [s+ t, C, s+ l] ∧ TRUE;

indica que si existen A → BC y $B \overset{*}{\Rightarrow} \omega_s, \ldots \omega_{s+t-1}$ y $C \overset{*}{\Rightarrow} \omega_{s+t}, \ldots \omega_{s+l-1}$, entonces $A \overset{*}{\Rightarrow} \omega_s, \ldots \omega_{s+l-1}$

Una vez que se han revisado las extensiones de longitud uno se pueden revisar las extensiones de longitud dos y así sucesivamente. En la Figura 22 el ciclo abarcador es el ciclo de longitudes, de la más corta a la mayor. Los tres ciclos interiores examinan todas las posibilidades, de combinaciones, de las posiciones de inicio, de las separaciones de longitudes y de las reglas.

```
(Boleano) chart [1..n, 1..|N|, 1..n+1] := FALSE
Para todo el conjunto de reglas
        Inicializar  s
        Para cada regla del tipo A → ws
            chart [s, A, s+1] := TRUE;
Para la longitud l, de la más corta a la más larga
        Para cada una, incializar s
            Para cada una dividir la longitud t
              Para cada regla del tipo A → BC
                  chart [s, A, s+1] := chart [s, A, s+ l] ∨ chart [s, B, s+ t] ∧ chart [s+ t, C, s+ l] ∧
TRUE;

regresa chart [1, S, n+1];
```

Figura 22. Algoritmo de análisis sintáctico ascendente de *tabla*.

Los árboles se obtienen modificando el algoritmo de reconocimiento para llevar el registro de los apuntadores de retroceso para cada arco que se va produciendo.

Evaluación cuantitativa

En este modelo consideramos como características las categorías de POS de las palabras, las reglas en sí mismas y el peso de las reglas. Para asignar valores cuantitativos analizamos la posibilidad de considerar lo siguiente:

- Las características del modelo corresponden al tipo de reglas empleadas. En este modelo se numeran las reglas por orden alfabético. La salida está ordenada por la prioridad de las reglas, por las POS y por el orden

alfabético de las reglas. Así que las variantes no están agrupadas con algún criterio.

Para ordenarlas podríamos ordenarlas en cuanto a reglas que varían del tope hacia abajo de la estructura. Una vez teniendo este orden y mediante análisis previo de algoritmos de clasificación de árboles probar diferentes clasificaciones por características para hacer sobresalir grupos similares.

- Las características satisfechas las asociamos con los diferentes POS. La idea es que sobresalga una variante para cada POS diferente y no varias. Por ejemplo, la palabra *la* tiene las categorías: PPR, PPR_C, N, Det. Así que si en un grupo de variantes solamente hay diferencias por una regla que utiliza diferentes POS de una misma categoría superior puede marcarse una de las variantes con un peso mayor para hacerla sobresalir de las demás.

  Otra posibilidad más laboriosa es reducir las marcas de POS a grupos. En el ejemplo anterior serían el grupo del sustantivo y el grupo de determinantes. Así que una de las variantes con Det y una con una marca seleccionada del grupo, en este caso nominativo, tendrán un peso mayor. Otro camino es la asignación de usos más frecuentes. Se relaciona al ejemplo de la palabra *una* que tiene 3 formas de verbo, y cuyo uso es mucho más probable como determinante. La solución sería darle un peso menor como verbo.

- Mayor diferenciación entre opciones considera el peso por la prioridad de las reglas. La prioridad de las reglas se aplica en forma descendente dependiendo de la posibilidad de asignar una estructura sintáctica. Las reglas de mayor prioridad se aplican primero y tienen un peso cero. Algunas oraciones requieren la aplicación de reglas de menor prioridad, pero aún en este caso, se aplican reglas de diferentes pesos y asociadas a ellos se cuantifica la diferencia entre variantes.

Sin embargo, la información sola de categorías de POS no nos ayuda a asignar pesos que diferencien las variantes correctas. Emplear las reglas para diferenciar grupos implica el uso de métodos complejos para hacer una clasificación de árboles en base a la cuál se podrían asignar valores cuantitativos. El peso de las reglas se utiliza directamente en el método por lo que siempre se obtienen las variantes con menor peso en general, es decir, con mayor prioridad. Solamente cuando se utilizan prioridades menores se utilizan reglas con diferentes prioridades.

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

3.7 CONSIDERACIÓN DE LA PROXIMIDAD SEMÁNTICA

Para describir el conocimiento semántico de contexto local en la oración, nos basamos en la idea bastante extendida, de que en la mente humana los conceptos se encuentran relacionados entre sí formando una red. Crear una red semántica es una tarea de labor intensiva en extremo, y difícil de lograr aún a largo plazo. En esta investigación consideramos la red semántica que se está desarrollando a partir de la red FACTOTUM[37]. La idea de su desarrollo se presenta en [Gelbukh et al, 98]. La idea en que se basa su uso para resolver la desambiguación sintáctica es buscar un análisis más profundo basado en la semántica de contexto. Para resolver la ambigüedad sintáctica, los enlaces de palabras o de grupos de palabras se realizan determinando qué tan cercanos semánticamente están esas palabras o grupos de palabras.

La determinación de la proximidad semántica se basa en las características de la red semántica, que son: los conceptos, las relaciones, y las trayectorias. Describimos la proximidad semántica como un valor cuantitativo, esta idea también ha sido empleada por [Sekine *et al*, 92] y [Rigau *et al*, 97]. Para determinarla no solamente consideramos la longitud por el número de enlaces sino también un peso asignado de acuerdo al tipo de relación. La trayectoria misma representa un valor cualitativo. Las relaciones explícitas tienen un valor por sí mismas que refleja la importancia de la relación. Mientras una relación ES_UN (“es un tipo de”) indica una cercanía entre pares de palabras y en algunos casos probablemente puedan ser sustituibles una por otra, la relación PUEDE_TENER refleja un mucho menor grado de cercanía, es lejana.

Entonces, la proximidad en las relaciones inferidas depende de la clase de

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

relación que se involucra y de la longitud de la secuencia lógica. Por ejemplo, puede existir una secuencia larga de relaciones ES_UN y entonces su valor será grande. De lo que se infiere que no solamente la longitud es importante. Ahora, si consideramos la red semántica completa tanto *humano* como *perro* son *animales* y podría llegarse a una trayectoria donde *perro* PUEDE_TENER *amigo*, por lo que el tipo de relación, donde la transitividad únicamente se da en cierto grado, también debe tener un peso.

Así que la proximidad entre un par de palabras es un valor que depende de la longitud y del tipo de relación. Para nosotros depende de las siguientes asignaciones:

1. Un valor para cada tipo de relación. Por ejemplo: 1 para ES_UN, 10 para PUEDE_TENER.
2. Valores específicos para enlaces individuales. Por ejemplo: el enlace *cosa* ES_UN *objeto* tiene una longitud mayor, o un peso mayor, que *Ford* ES_UN *carro*.
3. Un valor mayor a relaciones implícitas.

La primera asignación contempla los valores mismos de las relaciones explícitas, es decir, su importancia. Algunos de los valores asignados: ES_UN:1, PARTE_DE: 5, PUEDE_TENER: 10, ES_USADO_PARA: 5, etc.

La segunda asignación pretende corregir el problema que se presenta conforme las relaciones están más cercanas al tope de la jerarquía. Por ejemplo, en la Figura 23: *carro* ES_UN *objeto* y *libros* ES_UN *objeto*, por lo que se obtiene que la trayectoria entre *carro* y *libros* no es larga. En esta jerarquía mientras más alejadas del tope, las palabras tienen más aspectos comunes.

Por ejemplo una red como FACTOTUM tiene los siguientes niveles antes de llegar a un concepto de movimiento:

1) P. Physical Universe

2) Material Phenomena

3) Simple Actions of Physical Objects

4) MOTION

5) MOTION WITH REFERENCE TO DIRECTION

por lo que los tres primeros niveles deben tener valores de longitud muy grande, por ejemplo: 100, 75, 50.

La tercera asignación considera la problemática de las inferencias. Por ejemplo *carro* ES_UN *objeto* y *objeto* TIENE_SUBTIPO *libros*. De esta forma, la trayectoria es corta a pesar de que no hay muchos aspectos comunes. Para resolver este problema se asigna un peso mayor a una relación implícita que a una explícita. La precisión se obtiene junto con la segunda asignación que hace mayor la longitud de *carro* ES_UN *objeto* que de *Ford* ES_UN *carro*. Otro caso similar se presenta entre *Ford* y *maya,* que por la segunda asignación adquiere una distancia mayor. Así que

para relaciones implícitas un valor doble del de una explícita sería adecuado.

Desambiguación sintáctica

En el empleo de la red semántica para la desambiguación sintáctica realmente se está incorporando la componente semántica faltante en el módulo de las reglas ponderadas. La estructura sintáctica en este modelo se toma de la salida producida en ese módulo de modelo. Algunas de las gramáticas más actuales, derivadas de las gramáticas generativas precisamente incorporan restricciones semánticas, como la HPSG que las considera en la entrada de cada lexema en el diccionario. Esto equivale a tener la red semántica interna de cada palabra con las ligas a las posibles palabras con las que puede relacionarse en cualquier oración en el diccionario, lo cual implica una labor intensiva en extremo.

En nuestro método, esas restricciones semánticas se buscan en la red y se definen a través de la proximidad semántica (por ejemplo, ver **¡Error! No se encuentra el origen de la referencia.**), que involucra la distancia menor entre pares de palabras y su valor asignado. La evaluación de la proximidad no nada más está relacionada con estos valores obtenidos de la red misma, como se mostró anteriormente, sino que es necesario considerar además el tipo sintáctico de la relación. No todas las trayectorias son aceptables en un contexto específico. En

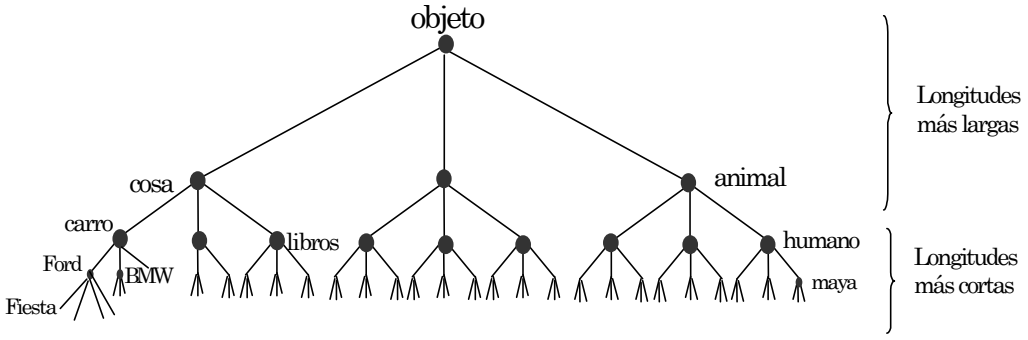


Figura 23. Diferentes longitudes en los enlaces de la jerarquía**.**

algunos casos se tendrá que buscar la trayectoria con las relaciones que sean más adecuadas al contexto sintáctico de la oración. Por ejemplo, si en la oración aparece la frase preposicional *con un telescopio*, la relación más cercana será USO y una relación más cercana tipo ES_UN no será la más adecuada para ese contexto.

Así que la tarea de desambiguación está muy relacionada con el método para encontrar las trayectorias aceptables mínimas y de contexto sintáctico. En una red

semántica existe un número infinito de trayectorias conectando dos palabras. Los aspectos matemáticos de solución y de implementación computacional se describen [Gelbukh, 98]. Para nuestro propósito solamente describiremos su uso para la desambiguación sintáctica.

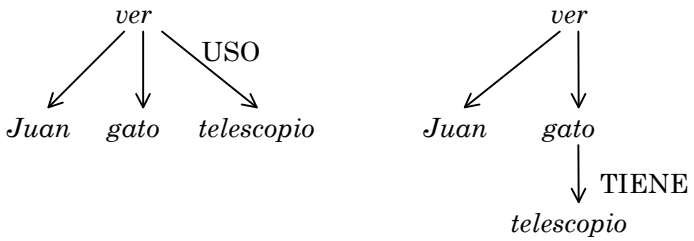


Figura 24 Ambigüedad sintáctica**.**

Consideramos el ejemplo de ambigüedad sintáctica, muy conocido, *Juan ve un gato con un telescopio*. El enlace de la frase preposicional *con un telescopio* puede hacerse a *Juan* o a *gato*. El significado entonces puede ser: Juan utiliza un telescopio para ver un gato, o Juan ve un gato que tiene un telescopio. Esta ambigüedad no puede resolverse con información léxica y sintáctica, únicamente, puesto que los enlaces son igualmente posibles desde esos puntos de vista. En la Figura 24 se muestran las estructuras posibles. Como se observa en esta figura, de acuerdo a los significados presentados, la primera estructura muestra la relación USO (*Juan usa un telescopio*) y la segunda estructura una relación TIENE (*un gato tiene un telescopio*). Por lo que una relación de ES_UN no es útil para desambiguar esta frase.

Las relaciones sintácticas cruciales para desambiguar la frase son: *ver* $\rightarrow$ *telescopio* y *gato* $\rightarrow$ *telescopio*. En la red semántica existe una trayectoria corta entre *ver* y *telescopio* en el fragmento de la red semántica para la frase *Juan ve un gato con un telescopio,* como se muestra en la Figura 25 con el tipo ES_USADO_PARA. Este tipo de relación se reforzaría con la indicación de una relación sintáctica instrumental entre *ver* y *telescopio*. La relación entre *gato* y *telescopio* es mucho más larga, a través de las relaciones más cercanas al tope de la jerarquía por lo que su peso es mayor. Por lo que en base a esas trayectorias se escoge la primer variante. En el caso más simple, la medida cuantitativa de la proximidad semántica (el peso de la longitud de la trayectoria) se emplea para una comparación.

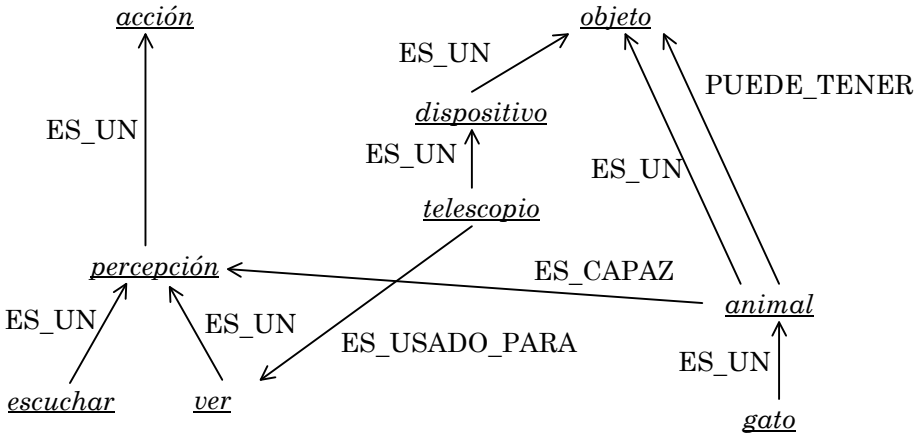


Figura 25 Red semántica para la frase, *Juan ve un gato con un telescopio*

Para una mejor calidad de análisis la trayectoria completa podría revisarse contra el tipo sintáctico esperado de la relación. Por ejemplo en la frase *Juan ve un gato con un niño*, existe una trayectoria corta entre *ver* y *niño* porque *niño* ES_CAPAZ *ver*. Pero en este caso, el tipo de relación contradice la hipótesis de que sea un instrumento para ver (*ve con un niño*). Esta es la razón que obliga a considerar todas las trayectorias posibles hasta que se encuentre una aceptable.

Evaluación cuantitativa

En este modelo nos basamos en las características del modelo que se emplean para obtener los valores de las trayectorias mínimas, es decir, para determinar la proximidad semántica entre palabras o grupos de palabras, consideramos lo siguiente:

- El número de características del modelo corresponde al valor obtenido de la proximidad semántica entre pares de palabras o grupos, normalizado. Como ya lo expusimos en la sección anterior, el valor total depende del valor de la relación, el valor del enlace dependiendo de su posición en la jerarquía y del valor por ser relaciones explícitas o implícitas.
- El número de características satisfechas indica las relaciones encontradas para la oración, en la red semántica, que concordaron con restricciones sintácticas a partir del modelo de reglas ponderadas

  Por ejemplo, *ver con un telescopio* tiene una restricción sintáctica marcada por la preposición *con* y asociada una restricción semántica de instrumento (en la misma red semántica). Lo anterior a diferencia del posible enlace *ver con un niño* de la frase *ver un gato con un niño.*

$$\text{Peso}_{PS} = (\text{Valor prox. semántica}) \times \left(\frac{\#\text{ restricciones satisfechas}}{\#\text{ restricciones planteadas}} \right)$$

3.8 ANÁLISIS SINTÁCTICO EN SU VERSIÓN ÚLTIMA

Cada uno de los modelos propuestos para el análisis sintáctico analiza las oraciones de entrada y obtiene diferentes variantes de estructura en la mayoría de los casos. La salida de cada módulo es el conjunto de variantes sin un peso asociado a su estructura sino en una secuencia de acuerdo a motivos arbitrarios del modelo mismo. Así que este "orden" se basa en características de construcción del método. Por ejemplo, en el modelo de reglas ponderadas, las estructuras de salida aparecen conforme a la secuencia de marcas de POS, al orden alfabético de las reglas aplicables y de acuerdo a la ponderación permitida de las reglas a aplicar.

Para la resolución de ambigüedad sintáctica, el nivel superior del analizador multimodelo realiza una estimación de las variantes para determinar cuales son las variantes sobresalientes. Primero requiere que las variantes de salida de cada módulo tengan un valor cuantitativo y en segunda requiere que sean compatibles. Proponemos que una asignación de pesos de acuerdo a características distintivas del método y a otra información estadística relevante disponible sea el valor cuantitativo. Por ejemplo, en el modelo de PMA, podemos considerar cuántos patrones de manejo avanzado se aplicaron para la oración analizada, qué probabilidad tiene cada realización de valencia, etc.

Modificamos la salida del modelo de reglas ponderadas a árboles de dependencias, de la misma forma que las estructuras de salida del modelo de PMA. De esta forma es posible considerar un método de evaluación que tome directamente las salidas de ambos módulos. Las variantes que se pueden obtener con el módulo de proximidad semántica también se modificarían a la misma estructura de dependencias, con el algoritmo de la sección 3.5. De hecho, la única diferencia entre las salidas modificadas de los modelos de reglas ponderadas y proximidad semántica es el peso asignado, como se verá más adelante.

La evaluación de las variantes sobresalientes (ver Figura 26) entonces puede

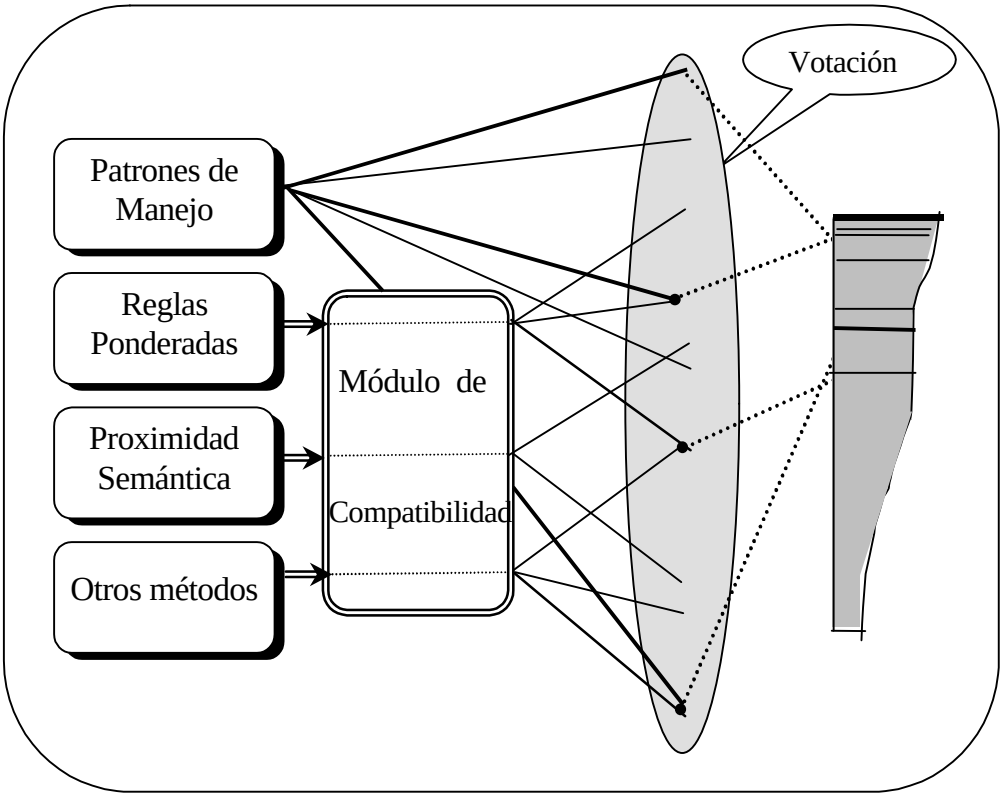


Figura 26. Modelo de análisis sintáctico y desambiguación

realizarse con un método de votación, en nuestro caso un modelo simple de votación. Primero sumando los valores cuantitativos de las variantes sin importar de que módulo salieron y enseguida ordenando las variantes por su peso. Esto permite que los diferentes conocimientos contribuyan con sus valores a las variantes. El módulo de votación del analizador selecciona las variantes con mayores pesos de entre todas las salidas disponibles. En la Figura 26 marcamos con líneas más gruesas las variantes con pesos mayores y las intersecciones indican las variantes con estructuras iguales.

Para poder hacer la votación entonces, nuestro modelo requiere una evaluación cuantitativa, para ordenar las variantes construidas por cada modelo, y una forma que las haga compatibles para su evaluación.

Ejemplos de evaluación cuantitativa

En las secciones de los modelos de reglas ponderadas y de proximidad semántica describimos la evaluación cuantitativa de las variantes. A continuación presentamos la evaluación en el modelo de PMA. En este modelo, mediante unas reglas, se emplean los patrones de manejo disponibles para las palabras que

constituyen la oración, con la finalidad de construir las variantes de análisis sintáctico. Las palabras que se buscan corresponden a los verbos, adjetivos y sustantivos de la oración que se analiza. Por lo que la especificidad del método se relaciona con el número de PM que se aplicaron en la variante, las valencias sintácticas empatadas, el número de homónimos (número de posibilidades de empate con cada palabra).

Así que para obtener una medida cuantitativa de la posibilidad de que una variante dada sea la correcta dentro de este modelo, consideramos las siguientes características:

- El número de características del modelo implica el número de patrones que se emplearon para cada variante. Por ejemplo, para analizar la oración *me percaté de que el banco estaba lleno de niños,* supongamos que se cuenta con los PMA correspondientes a *percatarse* y *lleno,* es decir un PMA de un verbo y otro de un adjetivo, de un total de cinco palabras con posibilidad de tener PMA: *percaté, banco, estaba, lleno, niños.*
- El número de características satisfechas corresponde a cuantas valencias pueden empatarse. Por ejemplo teniendo el PMA de $acusar_1$ es posible analizar la frase *María acusó a su jefe de fraude*. Donde se pueden empatar las tres valencias del PMA. En cambio en la frase *María acusó a su jefe* sólo pueden empatarse dos de las tres valencias.
- Las estadísticas disponibles, que obtenemos en base al método de adquisición semiautomática que describimos en el siguiente capítulo, corresponden a la frecuencia de uso de las realizaciones sintácticas de las valencias. Por ejemplo, para analizar la frase *habló con el director del CIC,* contamos con las frecuencias de empleo de *director de,* y de *hablar de.* Con estos pesos la variante con *director de* tendrá un peso mayor a la variante que considera el segundo caso.

Así que en base a esta información proponemos la siguiente medida cuantitativa:

$$\text{Peso}_{PMA} = \left(\sum_{tipo} C \frac{\#\,\text{PMA}}{\#\,\text{palab.contenido}} \right) \times \left(\frac{\#\,\text{Valencias empatadas}}{\#\,\text{total de valencias}} \right) \times \left(\prod \text{peso de } rsv \right)$$

Donde *C* es una constante que depende del tipo, si es verbo su valor es mayor, comparado con adjetivos y sustantivos; *rsv* significa realización sintáctica específica de valencias

A continuación presentamos en una forma breve la contribución de los modelos propuestos. Las características de las herramientas las detallamos en el siguiente capítulo.

Para la frase *El productor trasladó la filmación de los estudios al estadio universitario*, consideramos lo siguiente:

1. Patrones de manejo

   Consideramos la siguiente información:

   4.34896 trasladar, dobj_suj, obj:a, obj:de, x:?

   0.436967 trasladar, obj:a, x:?

   Donde los números de la primera columna representan los valores de realización sintáctica específica obtenidos del método de compilación de información para los patrones de manejo, método que se discute en el próximo capítulo. Los valores menores a uno corresponden a muy escasas apariciones, por lo que no los consideramos, como el:

   0.202283 estadio,pred:de

   La marca "x:?" representa una valencia repetida mediante clíticos.

   Así que considerando únicamente el patrón de *trasladar* se favorecen las variantes con la estructura de: trasladar algo a algún lugar desde otro lugar.

2. Con el modelo de reglas ponderadas obtenemos 8 variantes con el mismo peso.

3. Con la proximidad semántica, encontramos las siguientes relaciones

   filmación -> director

   -> subtipo de *espectáculo*

   trasladar -> con referencia a una dirección o a un lugar

   -> con relación a traslación de un objeto

   -> subtipo trayectoria

   estudio cinematográfico -> *l*ugar

   estadio -> como subtipo de *espectáculo*

   filmación -> cine -> director

   Así que únicamente la relación entre *trasladar* y *estudio* como "lugar" puede considerarse, con lo que favorece las estructuras *trasladar del estudio*.

En este ejemplo el método de patrones de manejo es el que da mayor contribución para reconocer las variantes correctas, sobresalen por los pesos de PMA principalmente y enseguida con la información del modelo de proximidad semántica.

Con la frase *Trasladó el productor la filmación del cortometraje al estadio universitario de la ciudad* aparecen otras consideraciones.

1. Patrones de manejo. La misma información anterior.

   En este ejemplo, la frase preposicional *del cortometraje* puede considerarse como la realización sintáctica “de algún lugar”, de igual forma la frase preposicional *de la ciudad* . Por lo que se favorecen estructuras ambiguas. A menos que hubiera un peso mayor para *filmación de* o para *estadio de*.

2. En el modelo de reglas ponderadas obtenemos 28 variantes con el mismo peso.
3. En la proximidad semántica contamos además de las anteriores con las siguientes relaciones

   filmación -> cortometraje

   ciudad ->gobierno- > edificio público -> estadio

De lo anterior, observamos que las frases preposicionales: *del cortometraje*, *de los estudios* y *de la ciudad* involucran mayor número de variantes en los tres modelos. Pero que con la proximidad semántica se puede enlazar la subfrase *estadio universitario de la ciudad*.

En este ejemplo el método de proximidad semántica contribuye con mayor información para la desambiguación.

Características de votación del analizador sintáctico

Los tres modelos emplean características específicas que permiten el análisis de las frases con diferentes conocimientos. Los patrones de manejo cuentan con cierta información léxica, sintáctica y semántica de la palabra misma pero desafortunadamente ni todas las palabras tienen PM distintivos ni contamos con los patrones para todas las palabras posibles. Las reglas ponderadas hacen uso de la categoría gramatical por lo que no distinguen palabras específicas pero presentan todas las posibles combinaciones conforme a las reglas gramaticales, y la proximidad semántica considera la semántica de contexto en la oración, que no está presente en los otros métodos

El método de votación que empleamos es un método simple de conteo. Existen otros métodos, el tipo regla o cuenta Borda es el más empleado, asigna puntos de una manera descendente a cada candidato para ordenar cuantitativamente las selecciones y después sumar esos puntos. En nuestro caso, cada modelo asigna esos puntos. De esa forma ordenamos las variantes, primero en base a la evaluación del método que las produce y después calculamos la suma de ellas para ordenar la totalidad de variantes, así que la variante ganadora es la de mayor valor. [Fishburn & Gehrlein, 76], [Van Newenhizen, 92] y [Saari, 94] han mostrado que el método de Borda es el método óptimo de votación posicional con respecto a distintas normas.

En el método Borda se intenta maximizar la satisfacción del votante. Asigna puntos de manera descendente a cada uno de los votos para la lista de candidatos con el fin de ordenar las selecciones. Obviamente la satisfacción del votante declina conforme los sucesos se van tomando de la lista de preferencias. Se asume como simplificación que la caída de la satisfacción sea igual entre cada rango. El resultado es que Borda otorga la elección a un grupo porque tiene más candidatos. Esto debido a que el método supone que los votos ABC prefieren B sobre C exactamente con la misma fuerza que prefieren A sobre B.

En nuestro modelo las preferencias se asignan conforme a características de cada método y no son iguales predefinidamente, salvo en el caso de las reglas ponderadas. Este método Borda privilegia la fuerza de preferencia, otro método, el de Tideman [Mueller, 96] no se basa en niveles de preferencia, privilegia cuántos prefieren A sobre B. Asignando diferentes puntos a diferentes rangos es posible tratar todos los tipos de métodos diferentes. Estos métodos posicionales [Black, 87] tienen el problema de que la elección se ve afectada por el número de candidatos que representan a cada facción.

De las características expuestas de cada método, en nuestro modelo, el número mayor de candidatos se obtiene del método de reglas ponderadas, donde además no asignamos una medida cuantitativa que las haga diferenciables. Por lo que la votación puede dar como resultado una clasificación que no sigue la ley Zipf como pretendemos, sino una salida sin salientes obvias, en los siguientes casos:

- En oraciones formadas de varias frases, donde escasos lexemas tienen asociados PMA.
- En oraciones donde los enlaces evaluados con proximidad semántica no tienen marcadas diferencias.
- En oraciones compuestas de varios verbos y pocos objetos para cada uno.

En estos casos, necesitamos entonces agregar alguna información para hacer mayor la distinción entre variantes clasificadas. Podemos seguir dos estrategias, la primera sería la aplicación de información más detallada en cada método, es decir mantener un método único de votación pero basado en una mayor complejidad de la asignación cuantitativa. La segunda estrategia es eliminar la votación y en su lugar insertar un módulo de multievaluación, como lo indicamos en la Figura 27. Emplear diferentes criterios de evaluación implica una selección de métodos para toma de decisiones multicriterios.

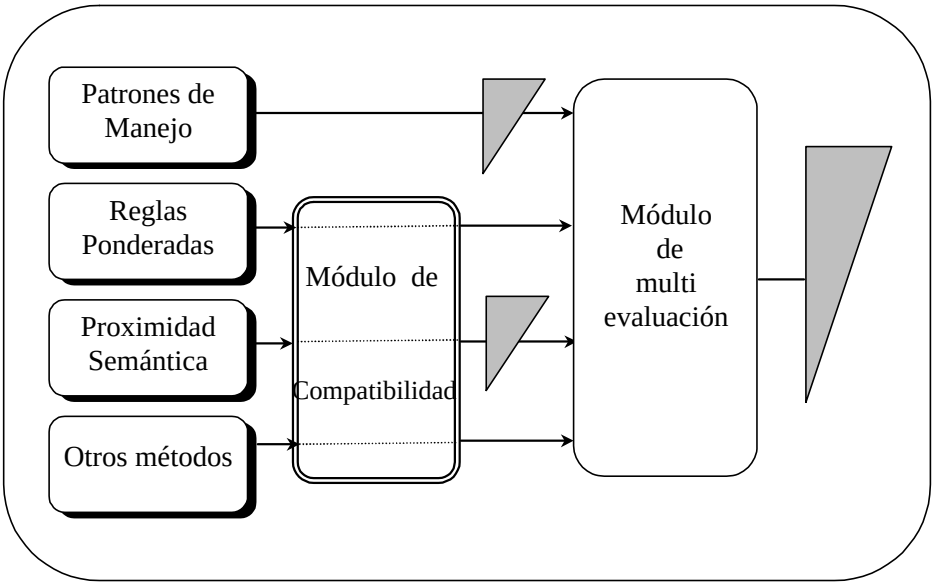


Figura 27 Multievaluación de variantes sintácticas.

[Lansdowne, 96] analiza diversos métodos de clasificación dadas múltiples alternativas y múltiples criterios, su objetivo es agregar información del criterio y obtener una clasificación total de alternativas. En este estudio, muestra que la presencia de empates en los criterios de clasificación pueden disminuir la potencia teórica de un método e incluso puede aumentar la dificultad de aplicarlo. También argumenta la necesidad de utilizar varios métodos de clasificación al mismo problema, para obtener diversas características que permitan una mejor decisión.

Para nuestro módulo de multievaluación podemos emplear diferentes métodos simultáneamente o en fases, entre los cuales podemos considerar: métodos de votación, métodos estadísticos, métodos lingüísticos o métodos híbridos. Muchos métodos, de los tipos mencionados, se han intentado como método único para desambiguar las variantes del análisis sintáctico, aunque su resultado no ha sido óptimo en esa tarea. Sin embargo en tareas más específicas como la que proponemos, de contribuir en la distinción de variantes ya clasificadas, su aplicación promete mejores resultados.

CAPÍTULO 4. COLECCIÓN DE ESTADÍSTICAS DE LAS COMBINACIONES DE SUBCATEGORIZACIÓN COMO MÉTODO PRÁCTICO

En este capítulo presentamos el método de obtención de los objetos de los verbos, de los sustantivos y de los adjetivos del español, es decir, el método de compilación del diccionario de patrones de manejo avanzados. Por la cantidad de entradas del diccionario, varios miles, no es posible compilarlo manualmente, más difícil aún es determinar las frecuencias o pesos requeridos, usando solamente la intuición lingüística de un hablante nativo

Cuando se dispone de un corpus de textos marcado sintácticamente y desambiguado, es decir, un corpus de textos con marcas de las relaciones sintácticas correctas, no es tan problemático calcular dichos pesos. Sin embargo, estas fuentes no existen para todos los lenguajes ni para todos los tipos de géneros de textos.

Por lo que proponemos un procedimiento semiautomático para compilar el diccionario de PMA a partir de un corpus de textos. Nuestro método tiene como objetivo primordial estimar los pesos de las combinaciones de los objetos de los lexemas predicativos, con los cuales se construirán los PMA del diccionario principal para la resolución de ambigüedad sintáctica.

En este capítulo, presentamos el algoritmo desarrollado, su aplicación a textos modelados para verificar su funcionamiento, y su aplicación a textos reales. Por último presentamos los resultados obtenidos en su aplicación al analizador básico.

4.1 MÉTODOS LEXICOGRÁFICOS TRADICIONALES DE COMPILACIÓN DE DICCIONARIOS EN OPOSICIÓN A LOS MÉTODOS AUTOMATIZADOS

Muchos de los requerimientos generales para definiciones lexicográficas son igualmente aceptables tanto para humanos como para dispositivos automatizados. Las decisiones lexicográficas impecables solamente las logran los lexicógrafos de alto nivel. Pero los especialistas en aplicaciones también logran decisiones de valor si éstas se basan en razonamiento, comparaciones, y experimentos de máquina.

Generalmente, los proyectos lexicográficos han requerido esfuerzos a muy largo plazo, y la participación de especialistas. Con la aparición de la computadora estos proyectos se han acelerado. Sin embargo, tienden a crecer en tamaño, por lo que el diseño y la construcción de diccionarios de varias decenas o centenas de miles de palabras es una tarea que involucra el trabajo de muchas personas durante años, en la especificación, el diseño, la compilación de datos léxicos, la estructuración de la información y el formateo adecuado para su presentación. Por ejemplo, la compilación del Diccionario del español usual en México [DEUM, 96] tomó varios años aún cuando emplearon algunas herramientas computacionales de la época, trabajo que se describe en [Lara & Ham, 79] y [García-Hidalgo, 79].

El uso del léxico implementado en computadora, lleva a una mayor convergencia de la teoría léxica y la práctica lexicográfica, ya que puede proveer información estadística y permite la manipulación de información en forma más rápida que además de facilitar el trabajo del lexicógrafo le permite hacer mejores decisiones. Por ejemplo, [Boguraev & Briscoe, 87] implementaron un algoritmo de transducción que toma los códigos gramaticales de LDOCE [Procter *et al*, 78] y produce códigos adecuados para otros formalismos gramaticales. El propósito del

proyecto fue convertir los códigos LDOCE para verbos que toman complementos al formato adecuado para el analizador sintáctico PATR-II [Shieber, 84]. Esta asociación clasifica los verbos en clases como sujeto de de ascensión, objeto equi, objeto de ascensión, etc.

Los métodos lexicográficos manuales requieren mucho esfuerzo económico y de tiempo. En la compilación de diccionarios por computadora pueden incluirse métodos lexicográficos que realicen algunas de las tareas de los expertos, para reducir el tiempo de análisis (oración por oración) de un corpus de textos. Por ejemplo, en forma de herramientas [Chodorow et al, 87] y en la automatización de tareas de compilación [Walker et al, 95].

Los métodos automatizados proporcionan estadísticas de experimentos que constituyen una herramienta muy poderosa y que no era accesible para los lexicógrafos clásicos. Estos métodos reducen el tiempo de trabajo de los expertos y permiten que una persona sin ser experto lexicógrafo pueda discriminar la información proporcionada para realizar las definiciones lexicográficas.

El método que proponemos para la compilación del diccionario de Patrones de manejo proporciona estadísticas de las combinaciones de subcategorización. Estas estadísticas representan las combinaciones que nuestro método selecciona y clasifica de acuerdo a su aparición en las variantes correctas del análisis sintáctico. Con esta información en una herramienta, los especialistas en aplicaciones realizan comparaciones que aunadas a su conocimiento lingüístico determinan las asociaciones de las combinaciones de subcategorización con sus respectivos actuantes.

4.2 INFORMACIÓN SINTÁCTICA PARA LOS PMA

La obtención de los patrones de manejo avanzados implica principalmente obtener el enlace de los grupos nominales y de los grupos preposicionales que realizan los objetos de los verbos, de los sustantivos y de los adjetivos. La fuente más común de ambigüedad sintáctica, y también el tipo de ambigüedad más difícil para resolver, es la ambigüedad al unir estos grupos.

Por ejemplo, para una frase simple como *Trasladaron la filmación desde los estudios hasta el estadio universitario*, puede asignársele al menos las interpretaciones sintácticas que se muestran en la Figura 28, donde mostramos los árboles de dependencias simplificados para cinco variantes. Realmente los árboles de dependencias son más detallados. Para este ejemplo, un hablante nativo escogería la primera estructura como la interpretación más probable, tomando en cuenta cierta información léxica.

Ahora, si consideramos únicamente POS, la estructura para el mismo ejemplo es V NP P NP P NP (V significa verbo, NP sustantivo o grupo nominal, y P preposición). Esta estructura de categorías sintácticas no proporciona toda la información necesaria para seleccionar la primera estructura para la frase del ejemplo. Podemos incluso dar ejemplos de frases para las cuales son correctas las otras cuatro estructuras. Por ejemplo, la variante 3 de la Figura 28, corresponde a la estructura correcta de una frase para el verbo *relatar*: *Relataron* [[*su vida desde los diez años*] *hasta su muerte*] que también tiene la estructura V NP P NP P NP.

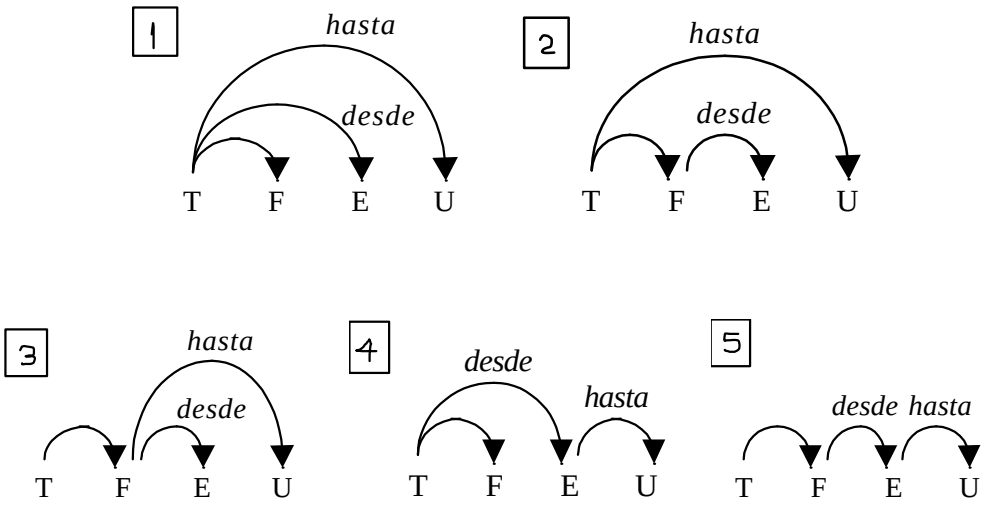


Figura 28. Variantes de la estructura sintáctica[38] para la frase
Trasladaron la filmación desde los estudios hasta el estadio universitario.

De lo anterior, se observa que la información léxica de cada palabra, relacionada al establecimiento de sus objetos, es la que ayuda a determinar la interpretación correcta. Por lo anterior, un analizador sintáctico necesita esa información léxica para desambiguar las frases, es decir, para eliminar las variantes incorrectas de la frase específica. Esta información no puede describirse mediante algoritmos o reglas, pero si puede obtenerse a partir de un corpus de textos.

Tanto para el análisis sintáctico como para reconocer los objetos de los verbos, necesitamos resolver la ambigüedad de enlace de grupos, especialmente, los preposicionales. Este problema se complica en el reconocimiento de los objetos de verbos, de sustantivos y de adjetivos porque deben ser los correspondientes a las realizaciones sintácticas de las valencias.

Para reforzar el objetivo de nuestro trabajo mencionamos que estudios cognitivos recientes [Schütze & Gibson, 99] sugieren que los seres humanos maximizan las relaciones de argumentos en la comprensión inicial de la ambigüedad objetivo, y que para describir esas relaciones se vuelven a considerar tratamientos combinados de léxico y frecuencia, además del conteo basado en lo último recordado. Consideran que se favorece el enlace de argumentos sobre los modificadores, los argumentos corresponden a las realizaciones sintácticas de las valencias y los

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

modificadores son los circunstanciales.

Dos líneas de investigación recientes, que parecieran ser adecuadas para la obtención de los patrones de manejo son: el enlace de frases preposiciones y la obtención de marcos de subcategorización. Estos estudios se han elaborado dentro del enfoque de constituyentes y pueden clasificarse en heurísticos, estadísticos, o basados en memoria. Algunos de ellos basados en corpus sin marcas, conocidos como *no supervisados*, y la mayoría basados en corpus marcados con la información que se pretende obtener, conocidos como métodos *supervisados*.

Trabajos relacionados: Enlace de frases preposicionales

Aunque para los patrones de manejo se requiere conocer las frases preposicionales que realizan las valencias sintácticas, no son adecuadas las aproximaciones desarrolladas bajo la línea de enlace de frases preposicionales. De los estudios iniciales, algunos se basan en [Ford *et al*, 82], que introdujeron la noción de *preferencias léxicas* para la resolución de ambigüedad. [Whittemore *et al*, 90] examinaron transcripciones de diálogos para intentar un algoritmo de enlace de frases preposicionales. [Hindle & Rooth, 93] describieron un método para aprender la asociación léxica a partir de un corpus de textos donde el objetivo son los patrones V NP P, y donde se asocia la preposición al verbo o al sustantivo. Estimaron estadísticamente la asociación léxica a partir de un corpus de entrenamiento, con marcas de POS y analizado sintácticamente. Notar que solamente es importante el enlace de la preposición y no las combinaciones completas como en la frase *dar un libro a Juan* donde se asocian al verbo *dar* los dos complementos (objeto directo e indirecto).

Otros trabajos como [Resnik & Hearst, 93] y [Ratnaparkhi *et al*, 94] consideraron la frase preposicional completa. Ambos emplean clases de palabras para determinar los enlaces, [Resnik & Hearst, 93] emplearon WordNet y [Ratnaparkhi *et al*, 94] obtuvieron las clases de palabras automáticamente mediante un procedimiento de clasificación de información mutua, basado en [Brown *et al*, 90]. En ambos trabajos se emplean métodos estadísticos de asociación para determinar los enlaces, en el primero entre las clases del elemento del lugar de enlace y del objeto de las preposiciones, en el segundo para obtener conjuntos óptimos de características (valores dependientes del grupo de 4 elementos V N1 P N2, y de las clases a las que pertenecen los núcleos de los elementos).

[Brill & Resnick, 94] describen un método de aprendizaje basado en transformaciones. Primero se enlazan todas las frases preposicionales al sustantivo y enseguida se comparan esas anotaciones contra las correctas del corpus. De esa comparación se determinan las transformaciones que se deben hacer para obtener los enlaces correctos. En cada iteración se intentan todas las transformaciones y se escogen las que resultan en mejoras generales. Estas últimas se añaden a una lista

ordenada de transformaciones y se aplican al texto, y así sucesivamente. En cada paso sucesivo se mejoran las transformaciones.

Los métodos basados en corpus, específicos para enlazar frases preposicionales, limitan su propósito [Yeh & Vilain, 98] al problema de enlazar frases preposicionales con un sustantivo o con un verbo, quizá debido a que en inglés las frases preposicionales típicamente ocurren al final de la oración, lo cual permite enlazarlas a los constituyentes precedentes. Esta es una simplificación que no resulta adecuada para nuestro propósito porque no considera el enlace: a adjetivos, a sustantivos y a verbos en un nivel más alto en la estructura jerárquica, o a oraciones completas (este caso fue considerado por [Chen & Chen, 96] para trabajos de traducción). Una desventaja de estos métodos es que las frases verbales e idiomáticas introducen seudo sustantivos que realmente no funcionan como puntos de enlace, ejemplo del primer caso: *pone atención a la clase*, el verbo es *poner atención*, ejemplo del segundo caso: *metió ruido en el convenio.*

Otra diferencia importante es que los trabajos antes descritos consideran el enlace de una sola frase preposicional. Una excepción es el trabajo de [Merlo *et al*, 97] que consideran los casos de enlaces de dos y tres frases preposicionales. El funcionamiento de su método para el enlace de tres frases preposicionales es de 43 por ciento para su conjunto de pruebas. Concluyen que el mayor problema en su método es la cantidad tan pequeña de casos con dos y tres grupos preposicionales, y por la posibilidad de demasiadas configuraciones.

Todos los trabajos mencionados usan corpus marcados sintácticamente, especialmente el Penn Tree-Bank Wall Street Journal [Marcus *et al*, 93], a excepción de [Ratnaparkhi, 98] que utiliza solamente un corpus con POS, aunque su método es para el patrón V N1 P N2.

Un trabajo para obtener patrones sintácticos es el de [Argamon *et al*, 98]. Ellos discrepan de la aproximación de detectar patrones sintácticos obteniendo el análisis sintáctico completo de una oración y extrayendo de ahí los patrones requeridos. Su objeción se basa en que en la mayoría de los casos es difícil obtener un análisis sintáctico completo para una oración además de que puede no ser necesario, en todos los casos, identificar la mayoría de los ejemplos de patrones sintácticos. Su estudio se basa en el análisis sintáctico parcial ([Abney, 91], [Greffenstette, 93]). Presentan una aproximación de aprendizaje general para reconocer patrones sintácticos en una frase. El método emplea un corpus de textos marcado con partes del habla, en el cuál todos los ejemplos de los patrones objetivo (los que se quieren obtener) se marcan sintácticamente con corchetes.

Todas las subcadenas de la frase de entrada se consideran como posibles patrones objetivo. El método calcula un puntaje para cada una de las subcadenas, comparándolas contra el corpus de entrenamiento. La comparación se realiza con evidencias positivas y negativas de cobertura de los patrones en el corpus de

entrenamiento. La frase de salida está marcada con corchetes de acuerdo con los patrones de mayor puntaje. Este método solamente reconoció los siguientes patrones S —V, V —O y secuencias de GN para el inglés, cuyo orden de palabras es más estricto que para el español. Notar que no se considera la información léxica puesto que sólo se refiere a categorías gramaticales.

Trabajos relacionados: Obtención de marcos de subcategorización

La otra línea de investigación, la obtención de marcos de subcategorización, se ha desarrollado manual y automáticamente para verbos principalmente. Entre los trabajos manuales más importantes están Alvey NL Tools [Boguraev *et al*, 87] y COMLEX [Grishman *et al*, 94]. En este último trabajo crean manualmente 92 marcos llamados *características de subcategorización*.

Debido a la utilidad de los marcos de subcategorización, trabajo reciente se ha enfocado a su estudio [Utsuro, 98] y a la extracción automática de esta información a partir de corpus de textos [Basili, 99]. Entre estos trabajos, [Brent, 91] inicia con un sistema que detecta cinco marcos de subcategorización a partir de un corpus sin marcas sintácticas. Su siguiente trabajo [Brent, 93] reconoce seis marcos y define un filtro estadístico para detectar los marcos verdaderos. [Ushioda *et al*, 93] describen un sistema similar en resultados pero introducen la obtención de estadísticas de los marcos.

[Manning, 93] describe la adquisición de un número pequeño de marcos de subcategorización para el inglés, [Monedero *et al*, 95] de uno aún más pequeño para el español. Manning, a diferencia de Brent prefiere métodos de detección menos fiables a expensas de obtener una mayor cantidad de marcos, obtiene dieciséis marcos de subcategorización, y se basa en filtros estadísticos para eliminar los errores. Este método no emplea herramientas muy desarrolladas, limita el espacio de búsqueda de cláusulas a las introducidas por la palabra *that* y por conjunciones, además del punto. Emplea un marcador de POS y como analizador sintáctico un autómata de estados finitos y un reconocedor de grupos nominales.

El trabajo más amplio es el de [Briscoe & Carroll, 97] donde describen un sistema capaz de distinguir 160 clases de subcategorización. También [Carroll & Rooth, 98] presentan una técnica de aprendizaje para obtener además de los marcos de subcategorización, su probabilidad de distribución para incorporarla a un analizador sintáctico. El sistema de Briscoe & Carroll emplea recursos muy desarrollados: un marcador de POS, un desambiguador de marcas de puntuación, un lematizador, un analizador sintáctico probabilístico, un extractor de patrones, un clasificador de clases de subcategorización, y estimaciones manuales a priori de esas clases basándose en corpus marcados sintácticamente. En este método, las posibilidades de subcategorización del verbo se definen en las reglas de la gramática y

directamente del análisis se toman los patrones de subcategorización. Existen 29 distintos tipos de patrones y 10 tipos diferentes para las frases preposicionales, que previa y manualmente se obtuvieron de otros diccionarios de subcategorización. La evaluación de los patrones se basa en el diccionario ANLT [Boguraev *et al*, 87].

Para lenguajes con un orden de palabras más libre y con un mayor número de preposiciones, la colección completa de marcos de subcategorización sería demasiado grande y muchas clases de subcategorización se requerirían para describir un verbo. Además de que en lenguajes con órdenes de palabras menos estrictos las realizaciones de los objetos también pueden ocurrir previos al verbo.

Así que estas líneas de investigación difieren en objetivo respecto a nuestra investigación, ya que nosotros requerimos una búsqueda exhaustiva y sin restricciones de todos los objetos para cada lexema predicativo. Cuando los objetos se realizan sintácticamente mediante frases preposicionales, en el español, éstas pueden ser más de una y enlazadas al mismo verbo. Como ya habíamos mencionado las preposiciones simples y compuestas en español incrementan los posibles marcos de subcategorización de un lexema específico por lo que de antemano no es posible definirlos en las reglas de la gramática sin perder la diversidad de composiciones que se presentan en el español. Requerimos entonces, un método de búsqueda exhaustiva de las valencias, considerando un orden de palabras más relajado que para el inglés y sin la necesidad de herramientas complejas.

Cabe mencionar que ni nuestro método ni los considerados previamente pueden tratar los casos donde el enlace difiere por razones semánticas y pragmáticas. Esos casos no pueden resolverse basándose en propiedades estructurales de la oración, por ejemplo: *Yo quiero ese carro en la foto.* En este sentido, los métodos basados únicamente en principios puramente pragmáticos también se equivocan en muchos casos. Los modelos basados en aproximaciones de Inteligencia Artificial de sentido común tienen diferentes problemas. [Jacobs *et al*, 91] indican que este tipo de modelos funciona bien en un dominio restringido y bien definido.

4.3 BASES DEL MÉTODO DE OBTENCIÓN Y EVALUACIÓN DE ESTADÍSTICAS DE OPCIONES DE ANÁLISIS SINTÁCTICO

El método que proponemos para obtener los objetos de los verbos, sustantivos y adjetivos del español también se basa en obtener las estadísticas de variantes del análisis, al igual que en el método de desambiguación sintáctica que describimos en el capítulo anterior, sólo que en este caso las variantes son las *combinaciones* de palabras individuales con preposiciones. Si nos basamos solamente en POS, estas combinaciones serían las componentes de los denominados marcos de subcategorización pero específicos para cada palabra y estas palabras pueden ser verbos, adjetivos y sustantivos. La selección de este tipo de combinaciones o marcos de subcategorización específicos, que en adelante sólo referiremos como *combinaciones*, no es aleatoria. Esas combinaciones son fijas, en un buen grado, para cada palabra específica, así que sus estadísticas son más confiables que las de palabras arbitrarias.

El diccionario que requerimos compilar es entonces una lista de posibles *combinaciones* (palabras con preposiciones) y, en el futuro, con algunas características de las palabras introducidas por estas preposiciones. En la forma más simple esa lista contiene entradas como las siguientes para *trasladar*:

1. *trasladar + hasta*
2. *trasladar + desde + hasta*
3. *trasladar + a*
4. etc.

Para resolver la ambigüedad de los enlaces nuestro método se basa tanto en las

frecuencias de las combinaciones en frases de textos particulares como en los errores del analizador sintáctico específico, es decir, en los árboles generados por el analizador sintáctico y en las estructuras que serían rechazadas ya sea por hablantes nativos o por otro tipo de procedimiento. En este método, para cada frase, se determina un *peso* (o probabilidad) para cada variante de estructura sintáctica. Este peso se basa en las estadísticas de las combinaciones en el lenguaje y en las estadísticas de variantes erróneas generadas por el analizador sintáctico específico.

Como ejemplo de este razonamiento presentamos un caso de desambiguación de POS. Supongamos que las frecuencias de diferentes POS en los textos bajo investigación son:

$$p^{+}_{sustantivo} = 0.4,\ p^{+}_{adjetivo} = 0.4,\ p^{+}_{verbo} = 0.2.$$

Cada variante consiste de solamente una característica: V_1 = {*adjetivo*}, V_2 = {*verbo*}, V_3 = {*sustantivo*}. Si esta es toda la información que tenemos, entonces dado el resultado del análisis $\boldsymbol{V}$ = {{*adjetivo*}, {*verbo*}} para una palabra dada, razonaríamos que puesto que

$$p^{+}_{adjetivo} > p^{+}_{verbo}$$

entonces la variante correcta debería ser *adjetivo*, ya que su peso es P ({*adjetivo*}) = 0.4 / (0.4 + 0.2) ≈ 0.66 mientras que el peso P ({*verbo*}) = 0.2 / (0.4 + 0.2) ≈ 0.33. En otro resultado tenemos que $\boldsymbol{V}$ = {{*sustantivo*}, {*adjetivo*}}, y entonces no puede hacerse ninguna decisión porque los pesos son iguales: P ({*sustantivo*}) = P ({*adjetivo*}) = 0.5.

Supongamos ahora, como usualmente sucede, que el marcador de POS empleado reporta a veces erróneamente algunas variantes para las palabras, y que lo hace con la frecuencia 0.9 para un sustantivo, con la frecuencia 0.1 para un adjetivo, y que nunca ha reportado un verbo erróneamente[39]. Entonces para el resultado $\boldsymbol{V}$ = {{*adjetivo*}, {*sustantivo*}} podemos decir que la respuesta correcta es *adjetivo* ya que ambos tienen la misma probabilidad y el analizador comete un error menor al marcar un *adjetivo*.

Entonces con este razonamiento, en nuestro método introducimos dos tipos de pesos estadísticos: p^{+} y p^{-}. El peso p^{+} significa la probabilidad, es decir, la frecuencia de ocurrencia de una combinación particular con la palabra rectora específica en el texto, en una estructura sintáctica correcta. Por ejemplo, en la Figura 28 la combinación *trasladar-desde-hasta* ocurre una vez en la estructura correcta.

El peso p^{-} es más interesante que el anterior, y hasta donde hemos investigado su uso no ha sido descrito en otros trabajos en el área, previamente; por lo que su

[39] La diferencia puede resultar de algún análisis de contexto que realiza.

introducción es un aporte teórico de esta investigación. Es la probabilidad de que la combinación ocurra en una estructura que fue construida por el analizador sintáctico, pero que no es correcta. Por ejemplo, en la Figura 28, la combinación *los estudios hasta* ocurre dos veces, en las variantes incorrectas 4 y 5; la combinación *hasta el estadio universitario* ocurre 1 vez en la variante correcta y 4 veces en las variantes incorrectas, entonces para esta última combinación, $p^+ = 1/5$ y $p^- = 4/5$.

El método de atribuir probabilidades a los objetos lingüísticos puede considerarse discutible. Primero, porque para cualquier intención semántica dada, no es de ninguna manera aleatorio el empleo de palabras específicas en un texto. Y en segundo, porque la acumulación de datos para distribuciones probabilísticas requiere mucho tiempo además de que no puede considerase universal debido a las particularidades de las fuentes. De hecho, los resultados son muy dependientes de la fuente [Roland & Jurafsky, 98].

Además, eliminamos la consideración de interdependencias ya que el incluir esos datos crea problemas en la implementación (de espacio, de tiempo, etc.). Su posibilidad de inclusión debe analizarse concienzudamente con la finalidad de decidir si vale la pena el esfuerzo de su consideración para los resultados esperados. Una alternativa también a futuro es que en lugar de tomar probabilidades, pudiéramos asignar algunos pesos apriorísticos [Briscoe & Carroll, 97] a las variantes y usar esos pesos en nuestros cálculos.

La característica de nuestro modelo de probabilidad es que el espacio de eventos se define en dos niveles de granularidad: léxica y sintáctica. El nivel léxico se relaciona a cada palabra y en el nivel sintáctico a los enlaces que forman parte de las combinaciones.

Deducción del modelo

Para elaborar las fórmulas de obtención de los pesos estadísticos de las diferentes variantes de árboles sintácticos de una frase, basadas en las combinaciones que aparecen en cada árbol, consideramos un modelo de generación de frases.

Consideramos que todas esas combinaciones que deseamos obtener aparecen en los árboles sintácticos de una frase como características abstractas del árbol. Numeramos esas características, por ejemplo, la combinación "*trasladar + desde + hasta*" es la característica número 1, "*acusar + a + de* " la número 2, etc. Denotamos esas características como f_1, f_2, etc. Entonces el conjunto completo o diccionario de estas características es $\boldsymbol{F}$.

Nuestro interés son las estadísticas de las características f_i (las combinaciones) y una simplificación en el modelo es omitir las relaciones entre ellas. Entonces consideramos una frase $\boldsymbol{P}$ como un conjunto de esas características, $\boldsymbol{P} = \{ f_{n_1}, \ldots, f_{n_K} \}$. Por ejemplo, para la frase *Trasladaron la filmación desde los estudios hasta el*

estadio universitario, obtenemos el conjunto ***P*** = {*trasladaron* + Ø + *desde* + *hasta*, *estudios*, *estadio universitario*}.

Para simplificar la discusión, suponemos que cada característica puede aparecer en una frase solamente una vez, ignorando las ocurrencias múltiples de la misma característica en una frase. Posteriormente indicaremos la forma de tratar este hecho. Consideramos también la elaboración del texto como un proceso de generación realizado por alguna fuente S, como un dispositivo que produce, una por una, frases $\boldsymbol{P}_m$.

El modelo de generación opera de la siguiente manera: para generar una frase ***P***, una fuente *S* conteniendo la característica $f_i \in \boldsymbol{F}$ decide si esta característica f_i será incluida o no en la frase ***P*** que se genera. La decisión se hace aleatoriamente, basándose en su probabilidad p_i, probabilidad que está asociada en el diccionario ***F*** a cada una de las características f_i. Por ejemplo, el generador *S* puede incluir la característica "*trasladar* + *desde* + *hasta*" en una de cada mil frases ***P***, con la correspondiente probabilidad $p_i = 0.001$.

Suponemos entonces que las características sí están incluidas, o no están incluidas en ***P*** de forma independiente. Obviamente, esto contradice la idea de coherencia en los textos como ya lo habíamos mencionado. Sin embargo, hacemos esta suposición, porque para este método no disponemos de un conocimiento léxico de colocaciones[40] [Smadja, 93], [Basili, 94], ni de atracción léxica[41] [Yuret, 98]. Además, en este método no pretendemos una cobertura total de ocurrencias concurrentes sino únicamente y de manera específica, de ocurrencias de combinaciones individuales. Así que basamos nuestro método en el único conjunto de datos disponible: las frecuencias p_i de combinaciones individuales $f_i \in \boldsymbol{F}$.

Basándonos en este conocimiento y en la hipótesis de independencia, podemos calcular la frecuencia de aparición de una frase específica ***P*** = { f_{n_1}, …, f_{n_K} } en la salida de *S* basándonos en las probabilidades de las combinaciones. Estas probabilidades se calculan de la siguiente manera:

$$a_n^{k,r} = \begin{cases} p_n^r \\ q_n^r \end{cases} . \qquad (1)$$

donde:

p es la probabilidad de que la combinación se seleccione

[40] Combinaciones recurrentes de palabras que ocurren concurrentemente más a menudo de lo esperado por casualidad.

[41] Es un modelo donde se asume que cada palabra depende de otra palabra en la oración, pero no necesariamente de una palabra adyacente.

q es la probabilidad de que la combinación no se seleccione y su valor es: $q = 1 - p$

k es el número de variante

r es 1 si corresponde a la variante correcta (se representa con "+").

es 0 si corresponde a variantes erróneas (se representa con "–").

n es el número de combinaciones

Así que tenemos las siguientes probabilidades:

p_n^+ si $f_n \in \boldsymbol{P}$ y k es la variante correcta

q_n^+ si $f_n \notin \boldsymbol{P}$ y k es la variante correcta

p_n^- si $f_n \in \boldsymbol{P}$ y k es variante errónea

q_n^- si $f_n \notin \boldsymbol{P}$ y k es variante errónea

Entonces la probabilidad de $\boldsymbol{P}$ es:

$$P(\boldsymbol{P}) = \prod a_n^{k,r} \qquad (2)$$

puesto que cada característica de manera independiente está incluida en la frase $\boldsymbol{P}$ con las probabilidades α. Donde r la denotamos como:

$$r = d_i^k \begin{cases} 1 & \text{si } k = i \quad \text{(variante correcta)} \\ 0 & \text{si } k \neq i \quad \text{(otras variantes)} \end{cases}$$

Estas probabilidades pueden verse como una matriz V con k filas, una fila para cada variante, y n columnas, una columna para cada combinación. Entonces los valores en la matriz son:

$$a_n^{k,r} = \begin{cases} p_n^{d_i^k} & V_k[n] > 0 \\ q_n^{d_i^k} & V_k[n] = 0 \end{cases} . \qquad (3)$$

donde $V_k[n]$ representan los valores de las probabilidades de ocurrencia de las combinaciones presentes, $n \in V_k$. Si la combinación n está presente, entonces $V_k[n] > 0$, de lo contrario $V_k[n] = 0$.

El diccionario $\boldsymbol{F}$ de las características es tan grande que cada característica específica f_n es mucho más frecuente que esté ausente en una frase a que esté presente. Este hecho es inherente a los textos, cada palabra, excepto algunas conjunciones y preposiciones muy comunes, aparecen en una minoría de las frases en el texto. En (2) el producto se toma para todas las variantes y para todas las combinaciones en $\boldsymbol{F}$.

Considerando el conjunto de las variantes de la estructura sintáctica

$\boldsymbol{V} = \{V_1, \ldots, V_N\}$ construidas por el analizador sintáctico para la frase $\boldsymbol{P}$, es posible usar la fórmula (2) para desambiguación. Supongamos que exactamente una de ellas es la correcta (de esta forma ignoramos los casos raros donde no se puede construir una variante correcta de estructura sintáctica para una frase dada). Sea H_j la hipótesis de que la variante V_j es la correcta. Sea ξ el evento de obtención de exactamente el conjunto $\boldsymbol{V}$, es decir, la matriz V, como el resultado del análisis sintáctico. Entonces, empleando la fórmula de Bayes tenemos:

$$\mathrm{P}(H_j \mid x) = \mathrm{P}(x \mid H_j)\frac{\mathrm{P}(H_j)}{\mathrm{P}(x)} \qquad (4)$$

Para abreviar, podemos denotar $\mathrm{P}(H_j \mid \xi) \equiv \mathrm{P}_j$, la probabilidad de que la variante V_j sea la verdadera. Puesto que exactamente una variante es verdadera se ve claramente que:

$$\sum_{\boldsymbol{V}_j \in \boldsymbol{V}} \mathrm{P}j = 1 \qquad (5)$$

Para calcular el valor de $\mathrm{P}(\xi \mid H_j)$ consideramos:

1º. No tenemos información a priori acerca de las probabilidades de las hipótesis individuales.

2º. Todas las variantes son ruido, excepto una que es la correcta.

Puesto que el evento ξ no depende de j por completo, podemos ignorar $\mathrm{P}(\xi)$, y como no tenemos información a priori acerca de las probabilidades de las hipótesis individuales[42], consideramos que todas tienen la misma probabilidad, tanto la correcta como las erróneas así que $\mathrm{P}(H_j)$ es una constante, entonces (4) puede reescribirse como (6):

$$\mathrm{P}_j \sim \mathrm{P}(\xi \mid H_j), \qquad (6)$$

donde ~ significa *proporcional*, es decir, $\mathrm{P}_j = C \times \mathrm{P}(\xi \mid H_j)$ con una constante de normalización C determinada de (5).

Si tuviéramos cualquier información a priori acerca de la probabilidad de variantes individuales V_i, por ejemplo basadas en la longitud media de enlaces sintácticos o en probabilidades de las reglas gramaticales correspondientes [Goodman, 98], o en algún otro parámetro, podríamos mantener el factor $\mathrm{P}(H_j)$ en (6).

Para calcular el valor de $\mathrm{P}(\xi \mid H_j)$ consideramos que todas las variantes son ruido, excepto una que es la correcta. Suponemos que la hipótesis H_j es verdadera, es decir, que $V_j = \boldsymbol{P}$ y todas las otras variantes V_k, donde $k \neq j$, son variantes espurias, es

[42] Que deberían basarse exactamente en los árboles V_i sin su comparación entre ellos.

decir, ruido. Supongamos que el ruido ocurre en el conjunto ***V*** independientemente de la estructura verdadera de ***P***. Nuevamente, esto no es correcto del todo, pero además de no tener ninguna información útil acerca de la naturaleza de su dependencia ni de sus interdependencias, es una simplificación común en los métodos estadísticos.

Entonces,

$$P(\xi \mid H_j) \sim \prod_{k=1}^{K} \prod_{n=1}^{N} a_n^{k,d_i^k} \qquad (7)$$

donde N es el número de combinaciones y K es el número de variantes.

Presentamos ahora una suposición sobre el ruido, considerando dos fuentes de información, una fuente de señales verdaderas que modela la variante verdadera, y una fuente de ruido, que modela todas las variantes incorrectas. Es decir, una fuente S^+ de frases correctas genera las frases ***P*** y una fuente S^- de errores genera las variantes de ruido del análisis (ver Figura 29), donde las figuras geométricas representan las combinaciones.

Así que el conjunto ***V*** de elementos (conjunto V_j de características f_i) es generado por ambas fuentes, S^+ y S^-. Solamente un elemento de ***V*** , la variante verdadera, es generada por la fuente S^+, y todas las otras por S^-. Entonces, un módulo de desambiguación recibe ese conjunto ***V***, y su tarea es estimar qué elemento de ***V*** generó la fuente S^+. La estimación se basa en las características f_i encontradas en cada uno de los elementos.

Supongamos que tenemos unas estadísticas de las frecuencias de características individuales en los elementos generados por S^+ y S^-. Entonces, considerando que la variante que se está generando es la correcta, S^+ incluye una característica f_i con la frecuencia p_i^+, y la fuente S^- la incluye con la frecuencia p_i^-. Nuevamente suponemos independencia, que las variantes generadas por la fuente S^- son independientes una de la otra y son independientes de la variante generada por S^+.

Partiendo de estas suposiciones, la hipótesis H_j es equivalente a la afirmación de que la variante V_j fue generada por la fuente S^+, mientras las restantes por S^-. La fuente S^+ genera las probabilidades p^+ y q^+, y la fuente S^- las probabilidades p^- y q^-.

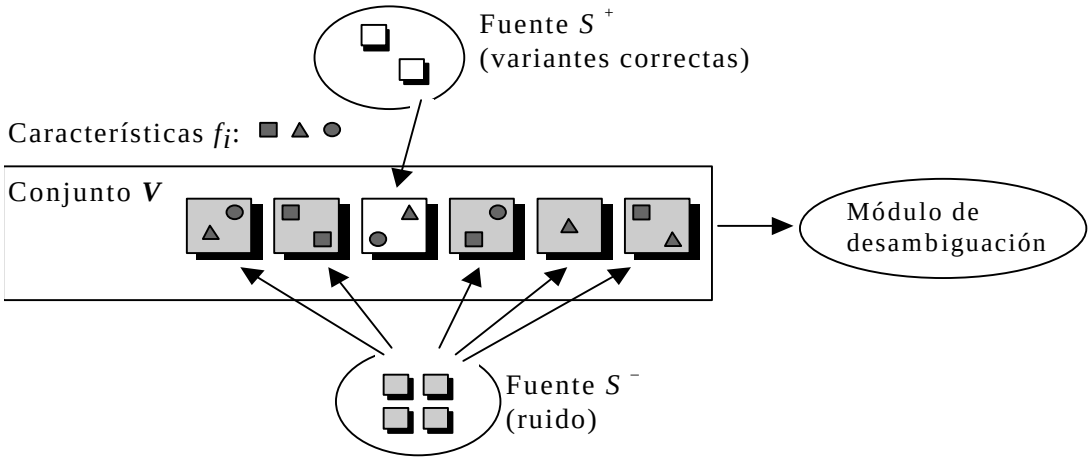


Figura 29 Modelo de dos fuentes de generación

Manipulando ahora algebraicamente la ecuación (7), mediante la introducción de un elemento unitario compuesto de las probabilidades q correspondientes a todas las combinaciones presentes en las variantes, es decir, para las p ($n \in V_k$) en toda la matriz, tenemos que:

$$\prod_{k=1}^{K}\prod_{n=1}^{N} a_n^{k,d_i^k} = \prod_{n=1}^{K}\left(\left(\prod_{n=1}^{N} a_n^{k,d_i^k}\right)\left(\frac{\prod_{n\in V_k} q_n^{k,d_i^k}}{\prod_{n\in V_k} q_n^{k,d_i^k}}\right)\right)$$

$$= \prod_{n=1}^{K}\left(\left(\prod_{n=1}^{N} q_n^{k,d_i^k}\right)\left(\prod_{n\in V_k}\frac{p_n^{k,d_i^k}}{q_n^{k,d_i^k}}\right)\right)$$

donde $\prod_{n=1}^{K}\prod_{n=1}^{N} q_n^{k,d_i^k}$ es la matriz llena de probabilidades q (de no selección de combinaciones). En esta matriz, las probabilidades q^+ están en la fila correspondiente a la variante correcta y las probabilidades q^- en K-1 filas. Ya que esta matriz no depende de las combinaciones presentes se puede eliminar. Esta manipulación puede verse como la limitación de la matriz a las combinaciones presentes en las variantes para la frase dada.

$$\prod_{k=1}^{K}\prod_{n=1}^{N} a_n^{k,d_i^k} = \prod_{n=1}^{K}\prod_{n\in V_k}\frac{p_n^{k,d_i^k}}{q_n^{k,d_i^k}} \qquad (8)$$

Nuevamente volvemos a manipular algebraicamente la fórmula anterior con el elemento unitario compuesto del cociente p^-/q^- para todas las combinaciones presentes en la variante correcta *i*.

$$\prod_{n=1}^{K} \prod_{n\in V_k} \frac{p_n^{k,d_i^k}}{q_n^{k,d_i^k}} = \prod_{n=1}^{K} \left(\left(\prod_{n\in V_k} \frac{p_n^{k,d_i^k}}{q_n^{k,d_i^k}} \right) \left(\frac{\prod_{n\in V_i} \frac{p_n^-}{q_n^-}}{\prod_{n\in Vi} \frac{p_n^-}{q_n^-}} \right) \right)$$

$$= \prod_{n=1}^{K} \left(\left(\prod_{n\in V_k} \frac{p_n^-}{q_n^-} \right) \left(\frac{\prod_{n\in V_i} \frac{p_n^+}{q_n^+}}{\prod_{n\in Vi} \frac{p_n^-}{q_n^-}} \right) \right)$$

Esta manipulación corresponde ahora a limitar el espacio de eventos a la parte de las combinaciones presentes en la variante correcta. El factor $\prod_{n=1}^{K} \prod_{n\in V_k} \frac{p_n^-}{q_n^-}$ corresponde a todas las combinaciones que no están presentes en la variante correcta, así que lo podemos eliminar con cierta pérdida:

$$\prod_{k=1}^{K} \prod_{n=1}^{N} a_n^{k,d_i^k} \approx \prod_{n\in V_i} \frac{p_n^+}{q_n^+} \frac{q_n^-}{p_n^-} = \prod_{n\in V_i} \frac{p_n^+}{p_n^-} \frac{(1-p_n^-)}{(1-p_n^+)} \qquad (9)$$

Como p^- y p^+ son valores pequeños, entonces $(1-p^-)/(1-p^+)$ tiende a uno, por lo que obtenemos finalmente:

$$\prod_{k=1}^{K} \prod_{n=1}^{N} a_n^{k,d_i^k} \approx \prod_{n\in V_i} \frac{p_n^+}{p_n^-} \qquad (10)$$

Así que para calcular el peso de la variante *j*-iésima, deben tomarse del diccionario ***F*** las frecuencias p_i^+ y p_i^- de todas las características f_i encontradas en esta variante V_j, y después aplicarse en la fórmula (10).

La crítica a los métodos estadísticos basados en corpus que se enfocan a las preferencias léxicas [Franz, 96], se refiere a las pocas variables estadísticas que se consideran, y que cuando se consideran varias de ellas, todo su manejo es en base a suposiciones de eventos independientes que no fueron motivados intuitivamente o que no tienen pruebas de que hay poca correlación. En este caso, la ecuación (10) es intuitivamente tan clara como que en la vida diaria, la gente cree en algunas noticias del radio y no cree en otras basándose en probabilidades de los eventos

correspondientes y la frecuencia en la cual las fuentes que las generaron cometen errores en un tipo específico de temas.

Limitaciones del modelo

En la ecuación (10) se presentan dos circunstancias a considerar, cuando $p_i^+ = 0$ y cuando $p_i^- = 0$. Cuando $p_i^- = 0$, la ecuación puede causar una división por cero. Este problema fue introducido artificialmente al manipular algebraicamente la fórmula (7) para obtener las ecuaciones (8) y (9).

Más adelante presentamos la forma de resolver el problema que se introduce con las combinaciones que aparecen escasamente y que producen que p_i^- sea muy bajo y p_i^+ / p_i^- alcance valores muy grandes.

La segunda consideración: $p_i^+ = 0$, está relacionada con el caso donde la frase contiene una palabra que no existía previamente en los datos de entrenamiento. Obviamente para cualquier combinación f_i conteniendo esa palabra, su $p_i^+ = 0$. Entonces $P_j = 0$ para toda j, lo que contradice la condición de normalización (5).

En lugar de introducir un caso especial para $p_i^+ = 0$, podemos usar un número muy pequeño, $\varepsilon << 1$, es decir, cuando $p_i^+ = 0$ lo cambiamos por $p_i^+ = \varepsilon$, y hacemos lo mismo con p_i^-. Esto no introduce una inexactitud significante y permite usar normalmente la expresión (10) en todos los casos.

La debilidad del razonamiento del modelo, y característica de muchos modelos estadísticos, son las hipótesis de independencia, principalmente la introducción de las combinaciones de las estructuras incorrectas de forma independiente a la estructura correcta de la frase, y la independencia entre combinaciones de una frase.

Sin embargo, estas dos hipótesis nos permiten usar las expresiones como $\prod_{V_j} p_i$ para las probabilidades de las variantes V_j, sin tomar en cuenta las dependencias entre las variantes V_k ni entre ellas y la frase ***P***. De otra forma, deberíamos tener datos cuantitativos útiles disponibles sobre esas dependencias. Los resultados obtenidos, que se discuten en la sección 4.6, no dan muestras de que esas hipótesis sean del todo erróneas.

Afinidades con otros métodos

Aunque no tenemos conocimiento de que otras investigaciones hayan empleado los errores del propio analizador para eliminar variantes incorrectas,

conocemos que en la teoría de radar, la detección de ataques se lleva a cabo por una evaluación diseñada cuidadosamente para medir las razones de falsa alarma de ataques recientes tanto como de razones de detección. Estas falsas alarmas son una indicación del radar de un objetivo detectado, aún cuando no exista ese objetivo, lo cual es causado por una señal de ruido o por niveles de interferencia que exceden el umbral de detección, que equivaldrían a la consideración de las variantes incorrectas generadas por el analizador sintáctico.

La teoría de señales, en la cual se basa la teoría de radar, se ha empleado en otras disciplinas, por ejemplo la sicología, en los años cincuentas y sesentas, como un intento de entender alguna de las características del comportamiento humano cuando se detectan estímulos muy tímidos, que no habían sido explicados por las teorías tradicionales de umbrales. En este caso se introduce un elemento de decisión, un acto cognitivo, para decidir si la señal está presente o ausente. Entonces, puede distinguir un éxito o un error cuando el estímulo está presente, o cuando el estímulo está ausente, la decisión será entre falsa alarma y rechazo correcto. En nuestro caso, la evaluación estadística p_i^- es equiparable a la falsa alarma.

En la teoría de radar, las evaluaciones iniciales de los sistemas de detección de intrusos tendieron a enfocarse exclusivamente a la probabilidad de detección, sin considerar la probabilidad de las falsas alarmas. Posteriormente al incluir sesiones de ataques previos en sesiones normales, se pudieron medir tanto razones de detección como de falsas alarmas, simultáneamente. Los conceptos matemáticos: espacios lineales de señales y proyecciones ortogonales, son conceptos claves para describir los problemas de procesamiento estadístico de señales como la detección y estimación [Franks, 69], [Scharf, 91], [Picinbono, 80].

En teoría de ecuaciones diferenciales de control automático y en otras disciplinas igualmente matemáticas, para probar la existencia de una solución se han empleado los Teoremas de punto fijo, uno de los más antiguos teoremas de este tipo es el de Brouwer [Debreu, 59] que es una generalización del corolario del Teorema de valor intermedio. Una generalización del teorema de Brouwer fue después simplificada por Kakutani [Debreu, 59].

El teorema de Brouwer establece lo siguiente: Sea $f : S \rightarrow S$ una función continua de un conjunto convexo, compacto, no vacío $S \subset R^n$ que mapea al mismo conjunto, entonces existe un $x^* \in S$ tal que $x^* = f(x^*)$ (es decir, x^* es un punto fijo de la función f). La prueba de la existencia de todas las soluciones es extremadamente difícil y no es posible en todos los casos.

Aunque en la teoría de señales, la incorporación de la medición de falsas alarmas tiene el propósito de solucionar la detección exitosa de los objetivos empleándola para regularizar las detecciones, en nuestro caso, la incorporamos para minimizar las variantes incorrectas, empleándola para medir la diversidad de variantes.

Proceso iterativo

Los pesos calculados con la ecuación (10) son los pesos de las variantes del análisis. Si tomamos una oración y la analizamos con un analizador sintáctico específico obtenemos un número de variantes de las cuales la mayoría son análisis incorrectos debidos al propio analizador. Si tuviéramos alguna información a priori para ese analizador acerca de las probabilidades de las combinaciones en las variantes, podríamos estimar las probabilidades de cada una de las variantes. Basándonos en estos pesos de variantes asignaríamos nuevos pesos a las combinaciones (p_i^+ y p_i^-) en la oración. Estos nuevos pesos de combinaciones permiten obtener a su vez nuevos pesos para cada una de las variantes de la frase, otra vez conforme a (10).

La ecuación (10) nos permite entonces obtener el peso de las variantes de análisis con los pesos anteriores de cada una de sus combinaciones y contribuir con esos nuevos valores para su reestimación. Teniendo unas probabilidades iniciales podemos proceder con todas las oraciones de un corpus, proceso que irá modificando tanto los pesos de las variantes como los pesos de las combinaciones. Conforme más oraciones se analicen, más datos contribuirán a los pesos de combinaciones y emergerán las combinaciones específicas, es decir, la información léxica. Este es el proceso iterativo que proponemos. Para obtener las estimaciones de todas las posibles combinaciones de un corpus y calcular sus pesos, desarrollamos el algoritmo que se detalla en la Figura 30.

El algoritmo resuelve dos problemas en los pasos iterativos:

- Resuelve la ambigüedad en el corpus a través de asignar pesos de probabilidad de las combinaciones a las variantes
- Compila el diccionario de las combinaciones para el diccionario de PMA y acumula los pesos estadísticos de las combinaciones.

El procedimiento iterativo se necesita para

1. Extraer las características estadísticas de ocurrencia concurrente de las preposiciones o de otras construcciones sintácticas con las palabras específicas, basándose en los resultados ambiguos de análisis sintáctico.
2. Usar los pesos estadísticos ya obtenidos para mejorar los resultados de análisis y resolver o disminuir la ambigüedad.

Estos dos pasos se ejecutan repetidamente. Al final, los pesos estadísticos de ocurrencia concurrente de diferentes combinaciones de las preposiciones con palabras específicas forman el diccionario de PMS, el cual es el modelo principal de nuestro modelo de análisis sintáctico del capítulo anterior.

El algoritmo termina cuando los pesos dejan de cambiar significativamente.

Conforme se obtiene la información de las ocurrencias de las combinaciones, en las variantes correctas y falsas, se acumulan los pesos de las combinaciones en un diccionario. Los dos conjuntos de pesos, los de variantes y los de las combinaciones, obtenidos después de un numero suficiente de iteraciones, son respectivamente el corpus con la ambigüedad resuelta y el diccionario base de PMA.

El conocimiento para definir las probabilidades a priori para el analizador sintáctico podría ser en base a longitudes de los vínculos, pesos de las reglas sintácticas, etc. Como por el momento no tenemos ningún conocimiento de este tipo que nos asegure un grado de corrección en las variantes, podemos hacer equiprobables las variantes del primer procesamiento del corpus.

1. En el inicio todos los pesos son cero
2. Para cada frase de entrada, se construyen todas las variantes de análisis de acuerdo a la gramática que el analizador sintáctico emplea.
3. Para cada variante se estima su peso w_k , conforme a (10), es decir, el producto de las frecuencias de las combinaciones presentes en la variante.
4. Los pesos se normalizan.
5. Cada variante se separa en estructuras locales de los nodos (ver Figura 31). Estas estructuras se incorporan al diccionario.
6. Para cada nodo de cada variante, se adiciona el peso de la variante al peso p^+, y el cálculo $(1 - w)$ al peso p^-.
7. Se toma nuevamente el corpus y se sigue al paso 3.

Figura 30 Algoritmo para calcular los pesos de combinaciones

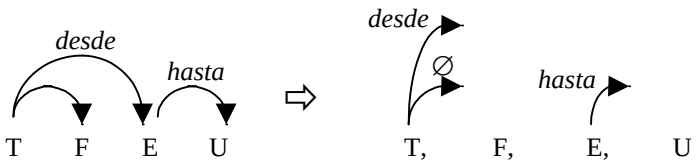


Figura 31. Las combinaciones como estructuras locales de los nodos para el ejemplo
Trasladaron la filmación desde los estudios hasta el estadio universitario.

La variante, de todas las estructuras generadas por el analizador sintáctico para la oración, con el peso más grande es la variante considerada como correcta. Al finalizar el proceso, el corpus con los pesos de las combinaciones es un corpus marcado sintácticamente. Con el procedimiento iterativo de reestimación propuesto podemos usar solamente textos sin marcas sintácticas.

4.4 CONVERSIÓN DEL MÉTODO EN SU APLICACIÓN A TEXTOS MODELADOS

Se han indicado algunos problemas cuando la fuente de datos de entrada es un corpus de textos. Algunas investigaciones han reportado que los procedimientos iterativos no son eficaces cuando el tamaño del corpus no es suficiente, como la de [Elworthy, 94]. Otras han encontrado diferencias significantes entre métodos experimentales y métodos basados en corpus, como la de [Roland & Jurafsky, 98] que llevaron a cabo pruebas de obtención de marcos de subcategorización sobre tres diferentes tipos de corpus. En ese estudio el método experimental se basa en completar oraciones o en crear ejemplos a partir de un verbo indicado. Así que por las dificultades de crear un corpus muy grande manualmente, y para evitar la dependencia de los resultados en un corpus específico, decidimos crear un corpus *artificial* para evaluar nuestro método.

Como una investigación inicial de las posibilidades del método para obtener los objetos de verbos, sustantivos y adjetivos, construimos una versión preliminar de programación del algoritmo. Esta versión funciona sobre el modelo artificial de texto. Construimos también un generador de estructuras sintácticas para poder hacer la prueba de efectividad del método estadístico, tanto de su exactitud como de su convergencia.

Para probar el método completo, construimos un prototipo del sistema simulador del proceso. En la Figura 32 mostramos el esquema general del sistema que consiste de dos módulos: el generador y el módulo del algoritmo. El generador abarca tanto la generación del diccionario como la generación del texto, así como la comparación de los resultados. El módulo del algoritmo da como salida las variantes ordenadas y las combinaciones del diccionario.

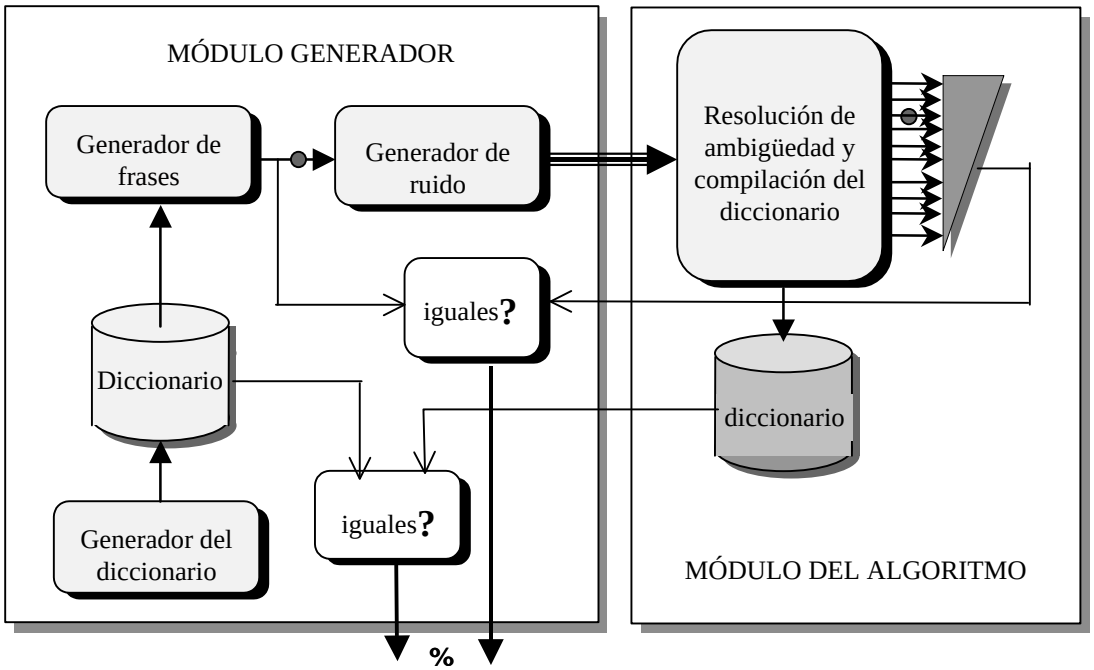


Figura 32. Esquema de prueba del algoritmo

Simulamos la generación de textos empleando solamente etiquetas numéricas en lugar de palabras. En la Figura 33 mostramos un ejemplo de una entrada del diccionario simulado. En nuestra implementación las frecuencias empleadas por el generador para el uso de palabras y preposiciones, para el número de actuantes de una palabra, para las combinaciones de actuantes, etc. corresponden con la ley de Zipf. Por lo que las frecuencias de las combinaciones en la figura son inversamente proporcionales a sus números de orden.

Generador. El generador produce frases aleatorias de acuerdo a las frecuencias y al diccionario programado en él. De esta forma, nosotros conocemos las frecuencias de todas las combinaciones de palabras y de preposiciones que el generador usa, pero el programa que implementa el algoritmo de compilación del diccionario las desconoce. Las frases generadas no son secuencias de palabras sino directamente las representaciones sintácticas de ellas. El conjunto de árboles sintácticos formales lo construimos mediante un programa simulador basado en un diccionario de PM, y este diccionario a su vez también fue construido mediante un programa simulador. La generación aleatoria, empieza por la longitud de la frase, después las palabras y las combinaciones.

Palabra874:	*Combinaciones posibles*:
3 actantes:	1. prep0 prep73
1: *obligatorio* prep4 prep0	2. prep4 prep12 prep0 3. prep0 prep31 4. prep4 prep7 prep31
2: *opcional* prep73 prep12 prep7	5. prep0 6. prep4 prep7 7. prep4 8. prep4 prep12 9. prep4 prep73 prep0
3: *opcional* prep0 prep31 prep7	10. prep4 prep7 11. prep0 prep0 12. prep4 prep73 13. prep0 prep12 prep31

Figura 33. Una entrada del diccionario simulado.

Después de que el generador produce una frase, genera adicionalmente en forma aleatoria variantes de su interpretación para modelar la generación de variantes incorrectas. Esas variantes incorrectas son el 20% de todas las posibles combinaciones del conjunto de palabras y del conjunto de preposiciones. De esta forma, la salida del generador contiene variantes de la estructura de la frase, algunas veces miles de ellas, pero sólo una de ellas es la correcta. Así que nosotros conocemos cual variante es la correcta pero el programa del algoritmo no la conoce. La variante correcta contiene solamente las combinaciones presentes en el diccionario del generador, mientras que las incorrectas pueden contener combinaciones que no están en el diccionario.

Algoritmo. Nuestro algoritmo analiza todos estos datos, es decir, la mezcla de la variante correcta con las variantes incorrectas. El algoritmo compila el diccionario, y también marca el corpus con las probabilidades de las variantes ya que estima cuáles variantes del análisis sintáctico son las correctas. Entre otra información, el diccionario contiene las probabilidades estimadas p_i^+.

Con este método, de antemano conocemos exactamente todas las propiedades del texto y del diccionario de combinaciones que empleamos para su construcción, y qué variantes son las verdaderas, por lo que al finalizar el trabajo, comparamos los resultados obtenidos con los valores *verdaderos* que fueron producidos por el generador. Consideramos como indicadores principales de la exactitud del procedimiento tanto el número de oraciones para las cuales se estimó la estructura correctamente, como el número de combinaciones del diccionario estimadas correctamente.

Este método sólo lo usamos para el desarrollo y depuración del algoritmo. Este algoritmo es del tipo *no supervisado*, es decir, no requiere una fuente de datos con los valores buscados previamente marcados, en cambio el método de simulación para su desarrollo y depuración sí es supervisado.

Experimentos

Este método permite experimentar con el programa de simulación, de tal manera que se pueden hacer los cálculos para diferentes corpus artificiales. Con el programa simulador realizamos una serie larga de experimentos, 53 iteraciones, con parámetros y procedimientos diferentes. Investigamos cuáles parámetros influían en la calidad de los resultados. Por ejemplo, encontramos que las combinaciones con baja frecuencia ocasionan problemas al método, pero también experimentalmente encontramos una forma de resolverlo.

Observamos que una causa significante de errores fue que algunas combinaciones aparecieron en muy pocas oraciones, por lo que el valor del cociente p_i^+ / p_i^- se incrementa y causa problemas. Éste fue un detalle significante encontrado empíricamente durante los experimentos. Encontramos que los resultados eran más estables y que la razón de error bajaba significativamente al suprimir los casos muy raros. Así que la solución fue añadir artificialmente algún ruido adicional a las características con frecuencia muy baja. Los mejores resultados se obtuvieron con la siguiente expresión:

$$\mathrm{P}_j \sim \prod_{f_i \in V_j} \frac{p_i^+}{p_i^- + l} \qquad (11)$$

En general, el método de simulación dio una base de investigación muy útil y cómoda. Los mejores resultados se obtuvieron cuando $l \approx S$, aunque los valores diez veces menores o mayores llevaron solamente a resultados peores de manera insignificativa.

Mientras mayor es el corpus mejor trabaja el procedimiento. En los experimentos, para un total de 1000 palabras, 100 preposiciones y 1000 oraciones, el número de variantes estimadas correctamente fue de 90%, y 87% para las frases ambiguas (que constituían el 75% del corpus). En promedio hubo 161 variantes por frase. En cambio para un corpus de solamente 200 frases la precisión fue de 85 y 80% respectivamente. La convergencia fue buena en ambos casos.

Los experimentos en los cuales se asignaron pesos iniciales de las variantes no dieron mejoras sustanciales en los resultados, comparados con los experimentos donde los pesos iniciales fueron todos iguales. Esto aún cuando los pesos iniciales distinguían bien entre variantes correctas e incorrectas, solamente incrementaron los resultados estimados iniciales de 18% a 37%.

Los resultados dieron una base de la efectividad del método. En los experimentos que llevamos a cabo, el porcentaje de las frases analizadas totalmente en forma correcta, o sea, de las frases para las cuales el árbol sintáctico completo estaba correcto, fue de hasta 87% de las frases ambiguas. Por lo que consideramos al método con muy buen grado de resolución de la ambigüedad.

4.5 CONVERSIÓN DEL MÉTODO EN SU APLICACIÓN A TEXTOS REALES

En la aplicación de nuestro método a corpus reales de textos aparecen complicaciones introducidas por parámetros que podíamos controlar en el ámbito de la prueba de nuestro algoritmo: el tamaño del corpus, diferentes géneros de textos, la estructura sintáctica, etc. En la aplicación del método a textos modelados estos parámetros se tenían controlados: el tamaño del corpus se podía variar así como la cantidad de palabras, la producción aleatoria en cierta medida podía representar los géneros, y dábamos como entrada las estructuras sintácticas de la oración en lugar de la oración misma. Entonces, la problemática de la aplicación de nuestro método a corpus reales de textos se divide en dos: la problemática del corpus y la del analizador sintáctico.

La aplicación a corpus reales de textos, implica la solución de diversos problemas del texto en sí mismo, es decir, el análisis de textos. La fuente de entrada debe analizarse respecto a varios criterios: la adquisición de los textos con la información requerida, la cobertura del corpus respecto a los fenómenos lingüísticos requeridos, la confiabilidad en el corpus, la independencia del método respecto al corpus, etc.

Idealmente es deseable emplear una muestra grande y representativa de lenguaje general. La razón de que la muestra sea grande es que mientras mayor es el corpus se espera un mayor número de palabras, lo que implicaría una mayor cobertura del diccionario del lenguaje, y principalmente supone mayor evidencia de los fenómenos lingüísticos diversos requeridos. Que sea representativa supone diversos niveles culturales del lenguaje, diversos temas y géneros. Sin embargo, estas cualidades no implican una a la otra, es más, en algunos casos se contraponen. Una contraposición que también se debe considerar es la que se presenta entre calidad y cantidad. El hecho de tener un corpus grande no garantiza que posea la calidad esperada.

Entonces, el corpus debería estar balanceado entre esas cualidades. Sin embargo, parece no ser posible balancear un corpus apropiadamente, al menos no sin un elevado esfuerzo. Además de que desafortunadamente los métodos de muestreo por ejemplo, para seleccionar calidad, son muy caros. Así que debemos asumir los problemas obvios de trabajar con datos desbalanceados, ya que construir un corpus balanceado requiere de mucho tiempo y de muy elevado costo.

Asumiendo la imposibilidad de tener un corpus con todas las cualidades deseables, podemos limitar las cualidades del corpus a las más importantes para nuestro método particular: la información requerida y el tamaño. Respecto a que el corpus tenga la información requerida, por ejemplo, [Biber, 93] indica el diferente uso de frases preposicionales según el género de los textos. [Roland & Jurafsky, 98] encontraron que hay diferencias significantes entre las frecuencias de subcategorización encontradas en diferentes corpus. Los autores identificaron dos fuentes distintas para esas diferencias: la influencia del discurso y la influencia semántica. La primera es causada por los cambios en las formas de lenguaje que se usan en diferentes tipos de discurso. La influencia semántica se basa en el contexto semántico del discurso. Por lo que un corpus con diferentes géneros sería muy adecuado.

Respecto al tamaño grande del corpus, los corpus actuales andan en el rango de un millón de palabras a cientos de millones, dependiendo del tipo, es decir, si son texto plano o con marcas de diversos clases. Por ejemplo [Berthouzoz & Merlo, 97] discuten que para obtener buenas aproximaciones de probabilidades, el corpus tiene que ser suficientemente grande para evitar los datos esparcidos y reflejar el uso natural del lenguaje. Ellas usaron el Wall Street Journal un corpus de un millón de palabras. A diferencia de ellas, en otros trabajos, no emplean el corpus completo sino subcorpus con características específicas para su investigación [Collins & Brooks, 95], [Yeh & Vilain, 98], [Ratnaparkhi, 98]. El corpus LEXESP que empleamos, tiene cinco millones de palabras, marcas de POS y diferentes géneros.

Para procesar el corpus, empleamos el analizador sintáctico de reglas ponderadas, ya descrito en el capítulo anterior. En cuanto a estructura sintáctica y número de variantes, [Church & Patil, 82] demostraron que para una gramática real, el conjunto de posibles análisis permitidos por la gramática para una entrada real analizada puede estar en los miles. Por ejemplo, en grupos nominales analizados usando una regla binaria recursiva ($N \rightarrow N\ N$) el número de análisis correlaciona con la serie de números Catalan. Un compuesto de 3 palabras tiene 2 análisis, uno de 4 tiene 5, uno de 5 tiene 14, uno de 9 tiene 1430, etc.

Este elevado número de variantes incrementa los pesos de combinaciones incorrectas, pero como nosotros los asociamos a las palabras específicas, el número de oraciones con esas palabras en el corpus y las combinaciones de adjetivos y sustantivos contribuyen a mejorar los pesos en las combinaciones correctas.

Para nuestro método la mayor importancia del corpus radica en la posibilidad de obtener las combinaciones para verbos, adjetivos y sustantivos. [Roland & Jurafsky, 98] explican que conforme la cantidad de contexto circundante aumenta (yendo de una sola oración a un discurso conectado) decrece la necesidad de expresar manifiestamente todos los argumentos del verbo. Esta situación también se presenta en las oraciones muy largas. Por lo que a diferencia de las oraciones para pruebas de analizadores sintácticos, para nosotros son de gran utilidad las frases que no son largas. Lo que nos permite eliminar las oraciones largas que presentan una cantidad elevada de variantes, más detalles se presentan en la sección 4.7.

Proceso general

El procedimiento que utilizamos es el proceso iterativo descrito en la sección 4.3. Como ya habíamos indicado, aproxima dos metas en pasos alternados: primero estima las variantes de análisis sintáctico basándose en los pesos existentes en el diccionario de combinaciones, después revalúa los pesos de sus combinaciones basándose en los nuevos pesos de las variantes de análisis sintáctico de cada frase, y

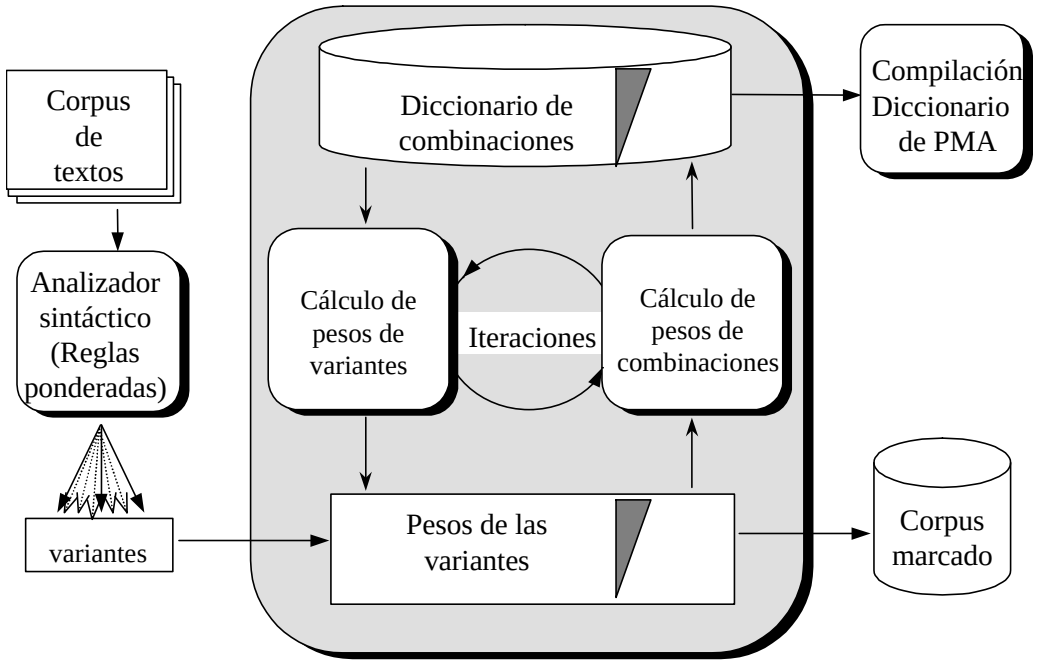


Figura 34. El procedimiento iterativo con corpus de textos[43].

[43] Con el triángulo mostramos un histograma de pesos, es decir, desde las variantes o

repite el proceso (ver Figura 34).

El proceso comienza con un diccionario de combinaciones vacío. En la primera iteración, para cada frase, todas las variantes producidas por el analizador sintáctico tienen los mismos pesos. Enseguida, se determinan las frecuencias p_i^+ y p_i^- para cada combinación encontrada al menos una vez en cualquiera de las variantes producidas por el analizador sintáctico para todas las frases del corpus.

Puesto que en esta etapa se desconoce cuáles variantes son las correctas, para determinar el número de ocurrencias de la combinación en las variantes correctas sumamos los pesos w_j de cada variante j donde se encontró la combinación, a p_i^+. Similarmente, para determinar p_i^- le sumamos el valor $(1-w_j)$ que representa la probabilidad de que la variante dada sea incorrecta. Entonces, podemos considerar todo el proceso de cálculo de los pesos como el proceso iterativo de solución de un solo sistema de ecuaciones, considerando la fórmula (9):

$$p_i^+ = \frac{\sum w_j}{S},$$

$$p_i^- = \frac{\sum (1 - w_j) + l}{V - S},$$

$$w_j = C \times \prod \frac{(p_i^+ + l)(q_i^- + l)}{(p_i^- + l)(q_i^+ + l)}$$

$$\sum w_k = 1$$

donde S es el número total de oraciones, V es el número total de variantes en el corpus. En las primeras dos líneas, la suma sólo se realiza para las variantes donde la combinación i aparece. El significado de l, como ya lo presentamos en la sección anterior, está relacionado con las palabras ausentes en el corpus.

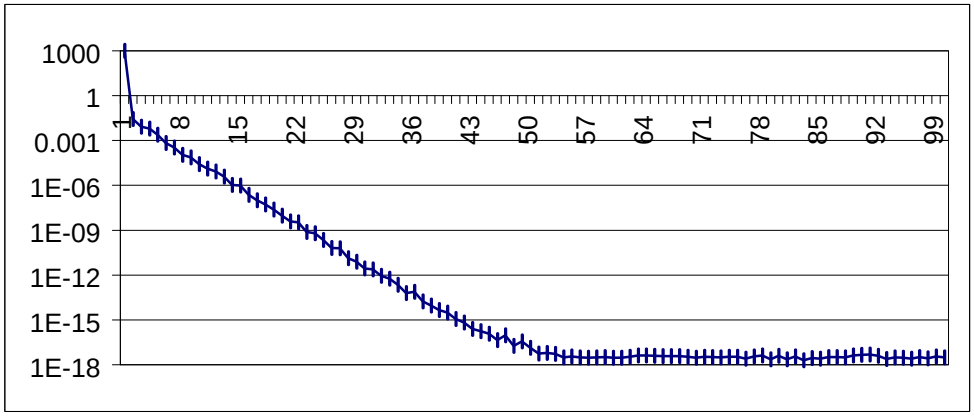
En la tercera línea, la multiplicación se hace para todas las combinaciones que aparecen en la variante j, para obtener su peso. En la cuarta línea la suma se hace para todas las variantes de la estructura de la frase específica bajo análisis, para normalizar. Los divisores en las dos primeras líneas y la constante C de la tercera línea solamente se introducen para normalización: S es el número total de variantes correctas supuestas y $(V-S)$ son las incorrectas. Así que los coeficientes los proporcionan el analizador sintáctico y el corpus de textos.

---

combinaciones con el mejor peso hasta las variantes o combinaciones con los peores pesos.

En los experimentos que realizamos, $\frac{S}{V-S}$ no probó ser la mejor opción, en su lugar experimentamos con diferentes valores: 0.01, 0.0001 y finalmente con 10^{-10} que probó ser la mejor opción. La expresión para λ es entonces un factor introducido como parámetro en las iteraciones, para alisar los efectos de los casos escasos, aunque es inherente al lenguaje que estos casos estén presentes.

Los valores de convergencia se muestran en la siguiente gráfica:



Realizamos también el cálculo de los pesos de las variantes con la fórmula (10) y no obtuvimos diferencias muy grandes en los resultados. Para este último caso la convergencia es muy similar a la anterior.

Al finalizar el proceso del corpus, tenemos el diccionario de todas las combinaciones encontradas, es decir, tenemos:

- Las frecuencias p_i^+ de las combinaciones en las frases *correctas*.
- Las frecuencias p_i^- de las combinaciones en las variantes *incorrectas* del análisis, es decir, en los errores del analizador sintáctico.

Teniendo estas probabilidades y un corpus resultante, analizado sintácticamente, a cada variante o hipótesis se le asigna un peso, el cuál es la probabilidad de que esa hipótesis sea la verdadera. El peso se calcula con (10) como un producto de los pesos en el diccionario de todas las combinaciones encontradas en él. Estos productos se normalizan dentro del conjunto de hipótesis producidas para la misma frase. Este proceso corresponde al siguiente algoritmo:

1. Para cada nodo del árbol de cada variante, se busca en el diccionario su estructura local, o sea, la *combinación*. Se calcula el peso w_i de la variante multiplicando los p^+ / p^- de cada combinación. Si no se encuentra la combinación en el diccionario, se usa una constante e pequeña.
2. Los pesos se normalizan mediante $\Sigma\, w_i = 1$.

3. Las variantes se ordenan por sus pesos. Las variantes con mayor peso se consideran como las correctas.

Pesos de las combinaciones y su uso

En el diccionario de combinaciones tenemos entonces sus pesos, basados en el corpus seleccionado. La utilidad de esos pesos se manifiesta en diferentes usos:

- Los pesos menores disminuyen el peso de la variante donde se encuentran. Por ejemplo, una combinación muy raramente empleada, incluida en el diccionario, no debe dar mucha preferencia a una hipótesis donde aparezca esa combinación.

- Los pesos de combinaciones desambiguan enlaces correctos pero opuestos. Por ejemplo una preposición puede introducir una valencia de un verbo y de un sustantivo en la misma frase. En este caso, la preferencia se da a las hipótesis conteniendo los enlaces más frecuentes, es decir, a la combinación que tiene un peso mayor. Por ejemplo "*Hablo con el director de la universidad*, donde *director de* tiene más peso que *hablar de*.

- Los pesos dan una idea de los errores que comete el analizador sintáctico. Un peso mayor a uno significa que la combinación aparece en variantes correctas, menor a uno que aparece en variantes falsas. Las combinaciones que tienen igual probabilidad en variantes correctas e incorrectas no ayudan, es decir, no contribuyen al peso de las variantes correctas y por lo tanto no es de valor alguno mantenerlos en el diccionario.

El cálculo de los pesos se rehace cada vez que una modificación significante se hace al algoritmo de análisis sintáctico o a la gramática.

4.6 EJEMPLOS DE VERBOS CON COMBINACIONES COMPILADAS AUTOMÁTICAMENTE

En el Anexo *Resultados del proceso iterativo*, se muestran los resultados al finalizar las 100 iteraciones del proceso del corpus LEXESP, que muestran las estadísticas de las combinaciones compiladas. Para analizar estos resultados escogimos una muestra de cuatro verbos, dos de ellos corresponden a ejemplos ya mencionados y dos más corresponden a nuevos ejemplos.

Las combinaciones se toman de los árboles tipo dependencias por lo que no hay un orden como en los árboles de constituyentes. Las combinaciones se denotan como:

Lexema, [*realizaciones sintácticas*$_{n}$]

donde las realizaciones sintácticas es una lista de dos tipos: frases preposicionales y grupos nominales. Las frases preposicionales solamente están representadas por la preposición. Cada grupo nominal está representado por "?". Hacemos notar que esta descripción no considera el orden, por ejemplo: *convertir*, *dobj_suj:?*, *dobj:?*, *obj*:*en* es equivalente a:

convertir, *dobj:?*, *obj*:*en*, *dobj_suj:?*

convertir, *obj*:*en*, *dobj_suj:?*,*dobj:?*

Donde *obj* indica objeto directo, *dobj_suj* indica sujeto pospuesto u objeto directo.

Esta representación es la forma más cómoda para obtener, revisar y aceptar las combinaciones correctas.

El siguiente ejemplo para el verbo *comprobar*, muestra la posibilidad de tener

el sujeto y el objeto directo pospuestos al verbo.

Combinación	p^+/p^-	p^+	p^-
Comprobar,dobj_suj:?,dobj_suj:?	3.26143	0.000179674	5.50906e-05
comprobar,dobj:?	2.19999	2.63402e-05	1.19729e-05
comprobar,dobj_suj:?	1.70269	0.000280344	0.000164648

Tanto en este ejemplo como en los siguientes solamente presentamos el rango de combinaciones con mayores pesos y omitimos todas las demás combinaciones para el mismo lexema pero con pesos bajos.

Para el verbo *acusar*, observamos que además de las valencias ya mencionadas a lo largo de los diferentes capítulos, se reporta la realización sintáctica de un circunstancial, representado por la frase preposicional introducida por "con".

Combinación	p^+/p^-	p^+	p^-
acusar,dobj_suj:?,obj:de,clit:?	11.0483	0.000293542	2.65689e-05
acusar,dobj_suj:?,obj:con,obj:de,clit:?	11.0483	0.000117417	1.06276e-05
acusar,dobj_suj:?,dobj_suj:?,obj:de,clit:?	11.0482	5.87084e-05	5.31382e-05
acusar,obj:con,obj:de,clit:?	11.0481	2.93542e-05	2.65693e-05
acusar,dobj_suj:?,dobj_suj:?,obj:con,obj:de,clit:?	11.0481	2.93542e-05	2.65693e-05
acusar ,obj:de,clit:?	4.3799	0.000335383	7.65731e-05
acusar,obj:de,obj:de,clit:?	3.53134	1.74375e-05	4.93792e-06
acusar,clit:?	2.94779	0.000383726	0.000130174
acusar,dobj_suj:?,dobj_suj:?,obj:a,obj:de	1.31443	1.6913e-05	1.28672e-05
acusar,dobj_suj:?,dobj_suj:?,obj:a	1.28471	3.96766e-05	3.08837e-05

También observamos que por su aparición en pocas oraciones, el sentido de

acusar como *revelar*, aparece únicamente con peso muy bajo: 0.157504 para *acusar,dobj_suj:?* y 0.152638 para *acusar,dobj_suj:?,dobj_suj:?*.

Tipos de elementos novedosos

Tomar las combinaciones del árbol de dependencias permite considerar todos los objetos del lexema, incluyendo los sujetos, y los objetos que en las oraciones se encuentran antes del verbo (realizadas como grupos nominales y clíticos). No se han considerado hasta ahora los clíticos que están insertados en el verbo, principalmente por el trabajo laborioso que requiere modificar el corpus LEXESP, sin embargo, este trabajo se considera a futuro.

Esta inclusión en nuestra investigación es muy importante, a diferencia de todos los trabajos de obtención de marcos de subcategorización ya mencionados en la sección 4.1. En esos trabajos no se considera el sujeto ni los objetos que se encuentran en un orden previo al verbo por el empleo de constituyentes, y también porque su objetivo es el lenguaje inglés, donde el orden de palabras es más estricto.

También se incluyó la consideración del sujeto pospuesto, que es más interesante que el sujeto previo al verbo, ya que en español siempre existe el sujeto, aunque no esté explícitamente se deduce de las características morfológicas del verbo.

Con este método, los valores bajos nos permiten dos tipos de acciones:

- A pesar de que las combinaciones correctas no cumplan con los valores esperados, mayores a uno, por la cantidad de oraciones con el lexema, es posible que los lingüistas, con una inspección visual rápida, reconozcan otras combinaciones correctas.
- Las combinaciones incorrectas presentan valores muy bajos e indican además de su ocurrencia en las variantes incorrectas, las frecuencias más usuales de sus POS.
- Las combinaciones incorrectas incluyen complementos circunstanciales obvios.

Por ejemplo para el adjetivo *ideal*, se obtuvo el peso 1.06275 para la combinación *ideal,pred:en,pred:para*, donde la frase preposicional circunstancial introducida por la preposición “para” es obvia como tal.

Ruido de información.

Existe cierta información que causa problemas en la obtención de las combinaciones. Además de problemas de datos escasos, marcas morfológicas y frecuencias de aparición, detectamos dos casos importantes:

1. Los objetos de los infinitivos cuya frecuencia es grande.

Por ejemplo, para el verbo *ir* que introduce infinitivos con la preposición *a*, obtuvimos la combinación *ir,dobj_suj:?,dobj_suj:?,obj:a,obj:a* con el peso 1.66206 debido a los infinitivos que a su vez realizan objetos con la misma preposición.

2. La preposición "de", cuya frecuencia es muy alta.

Por ejemplo la combinación *llenar,dobj_suj:?,dobj_suj:?,obj:de,obj:de* aparece con el peso 101.693 y la razón de duplicar un objeto con la preposición "de" se debe a su alto empleo en grupos nominales.

Como ejemplo del problema que se presenta por frecuencias de ocurrencia presentamos el caso de *comer*. Indica que aparece "*como*", con 3 POS: adverbio, conjunción y verbo, y que la aparición es mucho más frecuente con las dos primeras categorías gramaticales que con la tercera. Sólo la combinación *comer,clit:?* corresponde a frases del tipo *lo comieron*.

Combinación	p^+/p^-	p^+	p^-
comer,obj:de,obj:de,obj:en	7.54299	1.26722e-05	1.68e-06
comer,dobj_suj:?,obj:?,obj:sobre	4.20595	1.60786e-06	3.82283e-07
comer,dobj_suj:?,obj:de,x:?	3.4159	4.65501e-06	1.36275e-06
comer,dobj_suj:?,obj:sobre	3.04319	2.10207e-05	6.90744e-06
comer,dobj_suj:?,obj:a_punto_de	29.1575	0.000150785	5.17139e-06
comer,dobj_suj:?,dobj_suj:?,obj:de,obj:sobre	2.65452	9.06779e-07	3.41598e-07
comer,dobj_suj:?,obj:de,obj:sobre	2.49067	2.87282e-06	1.15343e-06
comer,obj:de,obj:en	2.30242	8.99539e-05	3.90693e-05
comer,clit:?	1.8792	0.000229021	0.000121871
comer,dobj_suj:?,dobj_suj:?,obj:sobre	1.80307	1.85296e-06	1.02767e-06
comer,obj:de,obj:sin	1.77636	3.04343e-07	1.7133e-07

La solución a este tipo de problemas sería seleccionar un grupo de oraciones donde se asegure el tipo de verbo deseado, o intentar corpus más grandes.

4.7 SINOPSIS DE ESTADÍSTICAS OBTENIDAS Y COMPARACIÓN DE TEXTOS MODELADOS Y REALES

Como ya habíamos presentado en la sección 4.5 la aplicación de nuestro método a corpus reales de textos se enfrenta a dos problemas: las características del corpus y los errores del analizador sintáctico. Así que los resultados obtenidos dependen principalmente de las características del corpus: tamaño y longitud promedio de oraciones, y de las características del analizador sintáctico: número de variantes de análisis por oración y oraciones no analizadas.

Las estadísticas de los corpus las presentamos a continuación:

	# palabras	**# preposiciones**	**# oraciones**	**# longitud promedio de oraciones**	**# variantes promedio**
Textos modelados	1000	100	1000	5 palabras	161
Textos Reales	~5 x 10^6	364	328,000	22 palabras	129 x 10^6

El corpus modelado fue 5 x 10^3 veces menor que el real en cuanto a oraciones, la longitud de oraciones fue de 5 palabras en el corpus modelado, es decir, 4 veces menor que la longitud promedio en el corpus de textos reales. Mientras mayor fue el corpus en textos modelados el procedimiento trabajó mejor. En textos reales, el procedimiento también trabajó mejor, de los resultados obtenidos en un corpus de 30,629 palabras y en el corpus completo de 5 millones de palabras, observamos que la

mejora fue de 25 veces en cuanto a resultados de combinaciones y en cuanto a definición de valores.

Mientras en el corpus modelado se conocían de antemano las variantes correctas de análisis sintáctico, en el corpus de textos reales se desconocen las variantes correctas y solamente pueden determinarse manualmente. El ruido introducido en el método de textos modelados, 20% de la cantidad total de combinaciones, es menor que las variantes erróneas obtenidas en el método de reglas ponderadas en promedio. Por lo que restringimos las oraciones analizadas a aquellas con un número máximo de variantes de 3000.

La cantidad de oraciones que no se pudieron procesar con textos modelados fue solamente de 4%. La razón fue solamente de trabajo, ya que se modeló el número máximo de variantes en función de la longitud de la oración y del número de preposiciones en ella, y si sobrepasaba las 50,000 variantes no se procesaba. En cambio en el corpus de textos reales la cantidad de oraciones que no se procesó fue de 40%, y las razones fueron las siguientes: errores en marcas de morfología, falta de marcas morfológicas (36% de las no analizadas), por exclamaciones y puntuación y por oraciones de lenguaje hablado. Donde exclamaciones y grupos nominales como oración son reproducciones de lenguaje hablado.

La convergencia del método fue buena, en los textos modelados se probaron hasta cincuenta iteraciones. En textos reales probamos para 100 iteraciones donde los valores se estabilizaban.

Los experimentos con asignación de pesos iniciales de las variantes no dieron mejores resultados que los obtenidos usando los pesos iniciales iguales en textos modelados, en los textos reales sólo probamos con algunas combinaciones manuales. Por la cantidad tan pequeña de combinaciones no es de considerar su resultado. Además de que no tenemos corpus marcados sintácticamente para obtener los pesos de esas combinaciones. Así que los PM compilados manualmente, nos sirvieron principalmente para otros propósitos que se describen en la siguiente sección.

Al experimentar con el programa simulado, teníamos la posibilidad de modelar diferentes “corpus”. Como mencionamos antes, el corpus que procesamos considera varios géneros, sin embargo no es un corpus balanceado. Por lo que trabajo futuro incluirá el proceso de corpus en dominios específicos, la compilación de corpus balanceados para pruebas y la obtención de corpus marcados sintácticamente.

4.8 COMPARACIÓN DE RESULTADOS DE LA OBTENCIÓN DE ESTRUCTURAS DE LAS VALENCIAS EN FORMA TRADICIONAL Y EN FORMA AUTOMATIZADA

La elaboración de los patrones de manejo que se han compilado para distintos lenguajes (inglés, [Mel'cuk & Pertsov, 87], ruso [Apresyan *et al*, 73], [Mel'cuk & Zholkovsky, 84], francés [Mel'cuk *et al*, 84, 88]) se ha realizado manualmente. Para el español compilamos, también manualmente un conjunto de patrones de manejo de cerca de 500 verbos [Galicia *et al*, 97], [Bolshakov *et al*, 98].

Esta compilación se realizó con la intención de cubrir el análisis de la mayoría de las peculiaridades de los PM de los verbos del español en su totalidad. Empleamos la descripción de manejo preposicional presentada en varias gramáticas y una colección de artículos políticos seleccionados de periódicos mexicanos actuales. Enseguida estos patrones fueron revisados por varios hablantes nativos. La colección cumplió con el objetivo de presentar una descripción teórica y descubrir peculiaridades imprescindibles en las descripciones de los PM.

De los manuales gramaticales consideramos el manejo de frases preposicionales. Como su estudio es general, analizamos estas descripciones para separar las frases preposicionales que realizan complementos circunstanciales de las que realizan los actuantes. El orden de palabras lo obtuvimos intuitivamente.

En este estudio detectamos la importancia de la descripción del uso de pronombres personales, como sujeto, objeto directo e indirecto, es decir, de pronombres personales con dirección. Esta descripción tiene dos aspectos:

- La descripción del sujeto pronominal. El sujeto pronominal siempre permanece en la forma de caso nominativo (yo, tú, él, etc.) y en el mismo lugar del sustantivo sustituido. Por lo que su descripción solamente considera la opcionalidad de realizarse mediante un grupo nominal o un pronombre personal.
- La descripción de objetos pronominales. Esta descripción es más complicada. Con las formas en imperativo, infinitivo y gerundio, estos objetos se expresan a través de formas clíticas (pronombres personales en los casos acusativo o dativo, o ambos), ligadas a la forma verbal. Por lo que se requiere la enumeración completa de los posibles órdenes del verbo con sus valencias expresadas de esta forma, y de su duplicación como ya se describió en la sección 2.6

Manualmente puede describirse de una manera general este tipo de particularidades pero la descripción de los órdenes usuales sólo se pueden obtener mediante el análisis de corpus de textos. Otro empleo del corpus de textos es la investigación adicional para reunir más estadísticas, que probaran el grado real de corrección de los PM compilados.

Con el método semiautomático, encontramos la diversidad en el uso de realizaciones sintácticas de las valencias y de sus posibles órdenes de aparición. Por ejemplo, encontramos que la modificación del orden, considerado normal en las oraciones del español, es muy usual. Esta situación es ignorada por algunos autores [Monedero *et al*, 95] por considerarse un recurso estilístico que aparece de forma excepcional.

También encontramos que la realización de valencias mediante clíticos y la duplicación de valencias tampoco deben despreciarse. Por todo lo anterior la estructura formal que presentamos en el Capítulo 2, la modificamos a la estructura final que presentamos en la Figura 35, donde incluimos la representación de las valencias duplicadas.

En la compilación semiautomática verificamos que algunas frases preposicionales que se marcan en los manuales gramaticales, corresponden a dominios muy específicos, y por lo tanto no se encontraron en textos comunes o su aparición fue extremadamente baja.

En la Figura 36 presentamos la estructura final formal obtenida en forma semiautomática para el verbo *acusar*$_1$.

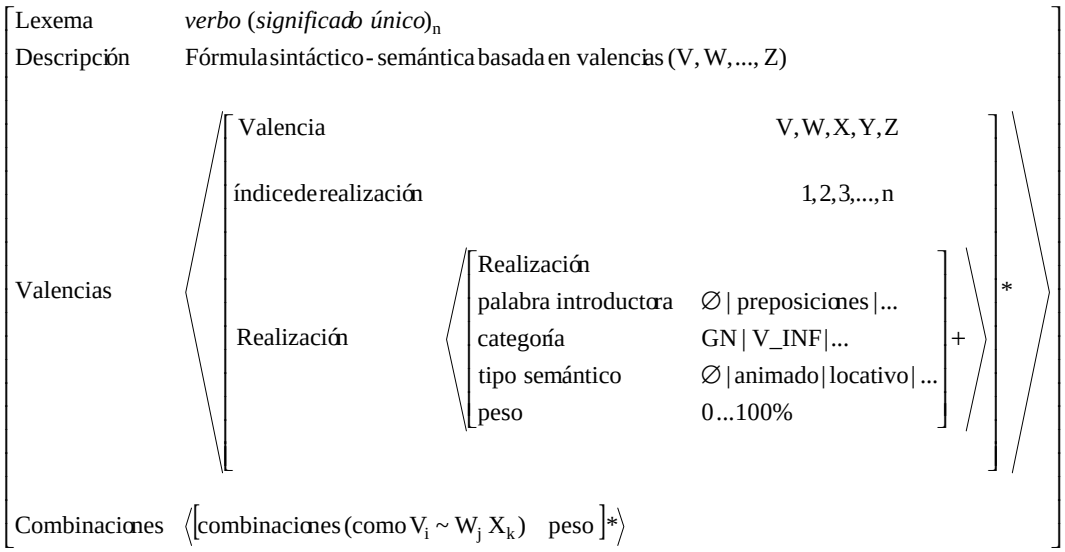


Figura 35. Estructura final formal de los PMA

La obtención de la estructura de valencias en la forma tradicional, es decir, en forma manual conforme a los métodos lexicográficos, se realiza mediante un trabajo introspectivo, buscando interiormente en la observación del propio uso del lenguaje los diferentes actuantes de los lexemas y sus realizaciones sintácticas. En la obtención en forma semiautomática, el trabajo que se realiza es confirmar la validez de las realizaciones sintácticas obtenidas y enlazar estas realizaciones con los actuantes de los lexemas asociados, es un trabajo más sencillo que va de las realizaciones sintácticas a los sentidos.

La idea del algoritmo para construir el diccionario de PMA con base en el diccionario de combinaciones es un algoritmo que incluye la participación de un lingüista. El algoritmo para compilar el diccionario es supervisado, a diferencia del algoritmo para compilar las combinaciones que es no supervisado. Brevemente, se describe a continuación.

Se presenta la lista de los lexemas para los cuales se obtuvieron las combinaciones del corpus y se escoge el lexema. Se extraen todas las combinaciones para el lexema dado. Se extrae la lista de preposiciones que aparecen en las combinaciones. Con esta lista se forman todos los grupos posibles, con la única restricción de que no aparezcan en el mismo grupo dos o más preposiciones que aparecen en una misma combinación.

De todos los posibles agrupamientos se eligen los que resultan en el mínimo número de conjuntos que contienen todas las preposiciones. Se ordenan los

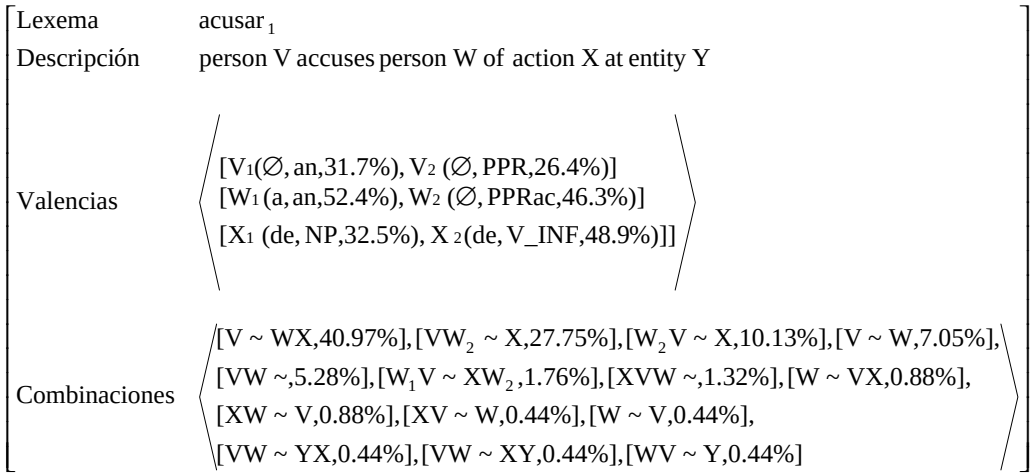


Figura 36. PMA para el verbo *acusar*$_1$

agrupamientos, primero los que empatan con las combinaciones encontradas de acuerdo a sus pesos. Los demás agrupamientos se ordenan alfabéticamente. Estos agrupamientos representan las realizaciones de cada valencia, es decir, un actuante.

Los agrupamientos se presentan al lingüista en ese orden, para su aceptación y asignación de la información sobre el significado de los actuantes. Después de que se hayan definido los actuantes, se solicita la confirmación de dos tipos de información:

- De actuantes obligatorios. El programa solamente los propone si aparecieron en todas las combinaciones.
- De hipótesis de incompatibilidad de los actuantes. El programa solamente los propone si nunca aparecen juntos en una oración.

Por ejemplo, para el lexema *huir* se tienen las combinaciones: *huir,dobj_suj:?,obj:hacia* con el peso 4.59491, *huir,obj:hacia* con 3.47915, *huir,obj:a* con 2.6778, *huir,dobj_suj:?,obj:a* con 1.27592, entonces un grupo de preposiciones es {hacia, a}

También debe ser posible que el lingüista agrupe manualmente algunos casos que no se encuentren. Otras facilidades necesarias para esta herramienta son la presentación de ejemplos y la posibilidad de presentar otras características de las realizaciones sintácticas de los actuantes, que deberían almacenarse previamente en el diccionario.

4.9 ALGUNAS CONCLUSIONES A FAVOR DE LA AUTOMATIZACIÓN

Ha habido una gran tradición de métodos empiristas, en contraste a los racionalistas, en la lexicografía. Para construir los diccionarios, estos métodos empiristas se basan en el análisis humano de hechos, es decir, en el análisis de textos reconocidos por su calidad y uso estándar del lenguaje. En estos estudios se define principalmente la información que desde el punto de vista de los hablantes nativos requiere explicación. Los diccionarios necesarios para el procesamiento lingüístico de textos, por computadora, tienen que detallar este conocimiento del lenguaje además del conocimiento que es obvio para los hablantes nativos y que no requiere descripción.

Las teorías gramaticales modernas actuales persiguen ese objetivo. Como ya lo habíamos mencionado, la MTM tiene este objetivo desde sus inicios. El diccionario concebido bajo la MTM, el diccionario combinatorio y explicativo, contiene las relaciones o correspondencias que se dan entre diferentes niveles del lenguaje, entre otras, en ella se definen las realizaciones sintácticas de las valencias y la correspondencia entre valencias sintácticas y actuantes. Razón por la cual nosotros hemos basado nuestra investigación en ella. Sin embargo, como [Kittredge, 2000] lo explica, la MTM requiere mucho detalle descriptivo y por lo tanto considerable tiempo para construirlo. Esto queda de manifiesto con el tiempo que Mel'cuk y sus seguidores han empleado en la compilación del diccionario combinatorio del francés [Mel'cuk *et al,* 84], [Mel'cuk *et al,* 88], [Polguère, 98].

Mel'cuk, mismo[44], considera que únicamente es posible desarrollar el diccionario combinatorio con lingüistas de habilidades intrínsecas, y con el estudio introspectivo (observación interna) de lingüistas entrenados para definir y describir

[44] En su participación de clausura de las conferencias CICLING-2000, realizadas en el CIC, IPN. México, D.F.

las formas del uso del lenguaje. Aunque reconoció que esta labor requiere meses para unos cuantos vocablos, contando con varios lingüistas especializados. Por lo que un diccionario de ese tipo aún a pesar de su gran utilidad es impensable a corto plazo, además de muy costoso.

Una forma menos larga de obtener los PMA pero costosa sería si se dispusiera de corpus de textos marcados sintácticamente. Este tipo de trabajo requiere mucho trabajo manual, aunque menor que el diccionario combinatorio. Sin embargo, se presentan problemas humanos, por ejemplo, [Leech & Garside, 91] discuten el problema que emerge al analizar sintácticamente de forma manual un corpus, relacionado a la disminución de exactitud y consistencia de los análisis con el paso del tiempo y del analista, la naturaleza de labor intensiva de producir análisis detallados, etc. Indican además que intentar que manualmente se construyan análisis consistentes con una gramática de cualquier tamaño y sofisticación pondría una enorme carga adicional en el analista.

Conforme han crecido las oportunidades de obtener corpus y marcarlos automáticamente ha sido más fácil compilar diccionarios tradicionales y para computadoras. Ahora es más común que los lexicógrafos empleen los corpus de textos, por las múltiples ventajas que ofrecen. Sinclair escribe en el prefacio de COBUILD [Sinclair *et al*, 87] que por primera vez un diccionario había sido compilado por la inspección detallada de un grupo representativo de textos en inglés, hablados y escritos, con millones de palabras. Que esto significaba que además de las herramientas para hacer diccionarios comunes (lectura y experiencia amplias en el inglés, otros diccionarios y por supuesto ojos y oídos) ese diccionario se basaba en evidencia física, mensurable. Recientemente, algunas de las mayores casas lexicográficas coleccionan grandes cantidades de corpus de datos.

Con nuestro método es posible compilar semiautomáticamente el diccionario de PMA, por supuesto sólo en la parte relacionada a combinaciones posibles e imposibles, no en la parte del significado de los actuantes, ya que el significado de los actuantes debe asignarse manualmente. De esta forma, se restringe la labor de un lingüista, ya que con toda esta información estadística solamente tiene que hacer las ligas de nivel superior.

Aunque no se obtiene la meta deseable de los métodos por computadora de eliminar la labor humana, un diccionario de este tipo ha requerido hasta ahora de la elaboración totalmente manual, además del tiempo y el costo, en gran cantidad. Una ventaja de que no sea automático es que algunas combinaciones, que no alcanzan a presentar valores distintivos porque no aparecen en el corpus con mayor frecuencia, pueden recuperarse por la selección del lingüista. Además de que el desarrollo en el área semántica no provee todavía los métodos requeridos para asignar la información semántica de los actuantes.

Así que la ventaja de la automatización es obtener resultados en un tiempo

mucho más corto, aunque no un resultado total. Otras ventajas son la rapidez con que puede trabajarse en distintos dominios y distintos corpus de textos. También la posibilidad de determinar el uso actual de las combinaciones en esos dominios, en distintos géneros y en variantes del lenguaje mismo.

4.10 REALIZACIÓN DEL SOFTWARE

El sistema que compila los pesos de las combinaciones consta de los siguientes módulos:

1. Un módulo que analiza sintácticamente el corpus de textos y obtiene para cada oración: el número de variantes, el número de combinaciones y las combinaciones de cada variante.

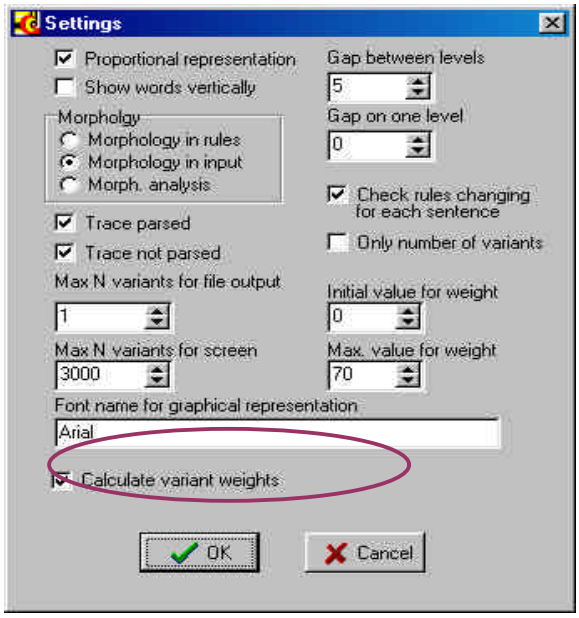
   Este módulo emplea el analizador básico con la gramática de reglas ponderadas. Con opciones para definir el número máximo de variantes, el rango de pesos en las reglas ponderadas, la longitud máxima de las oraciones, las relaciones de dependencia consideradas en las combinaciones, y otras opciones para facilidades de proceso.

2. Un módulo que realiza el proceso iterativo de calcular los pesos de las variantes, y los pesos de las combinaciones, alternadamente.

   Con opciones para seleccionar el tipo de cálculo o el valor de lambda, el cálculo de los pesos de las variantes (entre ellas las fórmulas 9 y 10) y opciones de facilidades de proceso.

3. Un módulo para crear una base de datos con las cadenas de las combinaciones y los pesos calculados para uso del analizador básico.

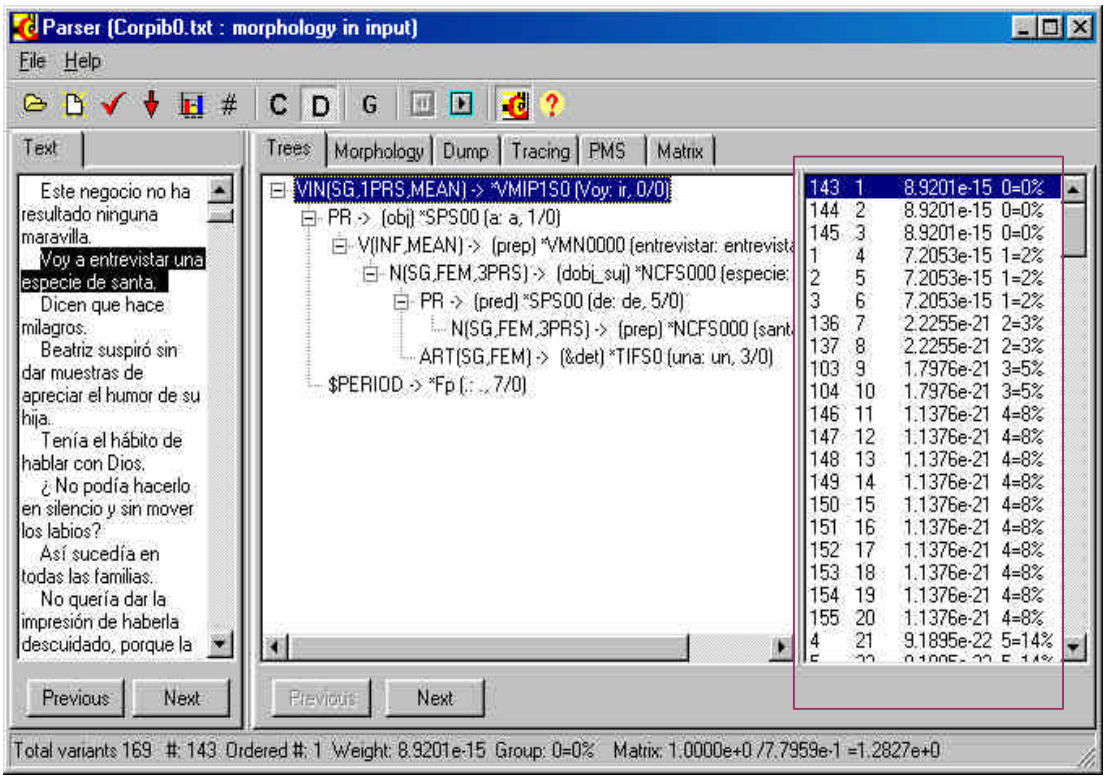
El analizador básico permite emplear los pesos de las combinaciones, como se muestra a continuación, con la marca en “Calculate variant weights”:



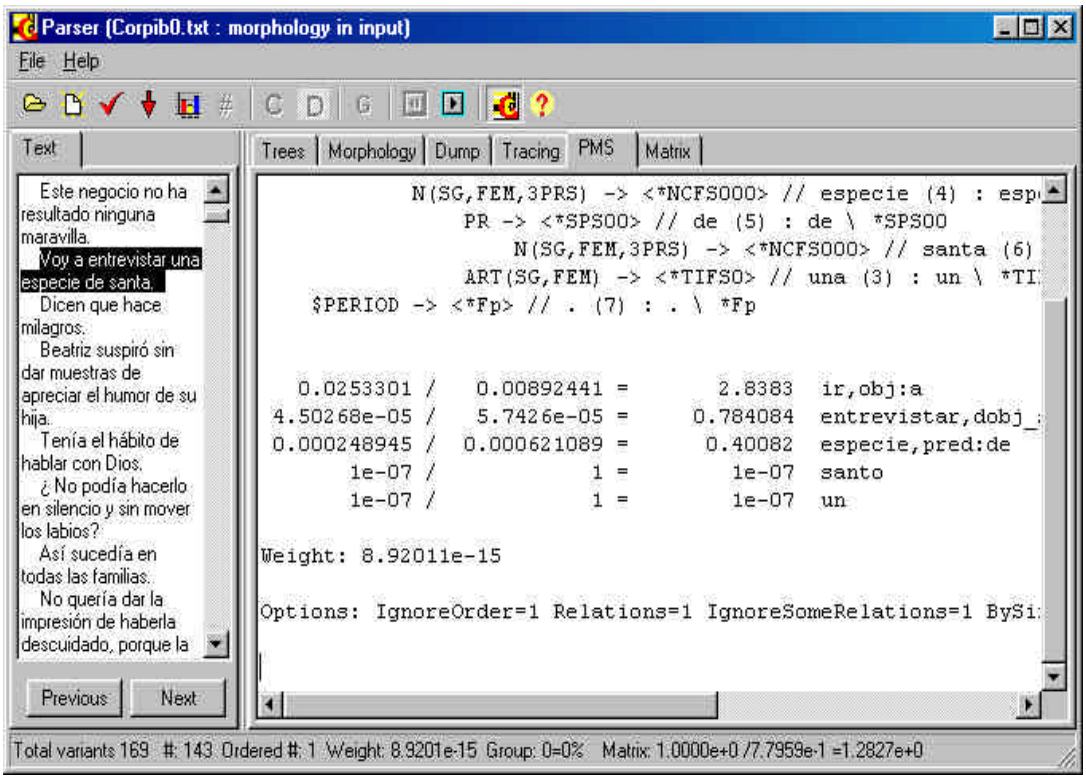
El proceso de compilación es tipo batch. Se realizó en máquinas Pentium II- 64K de memoria. Por el tiempo de proceso, el corpus se dividió en tres partes y se procesó paralelamente. El proceso de análisis sintáctico y extracción de combinaciones tomó 15 horas para cada tercera parte del corpus total. El proceso iterativo de cálculo de pesos de variantes y pesos de las combinaciones alternadamente, tomó 17 horas para las 100 iteraciones.

A continuación mostramos algunos resultados de la aplicación de los pesos obtenidos de las combinaciones en el analizador básico. Los resultados totales los presentamos en la sección siguiente, aquí sólo presentamos dos ejemplos.

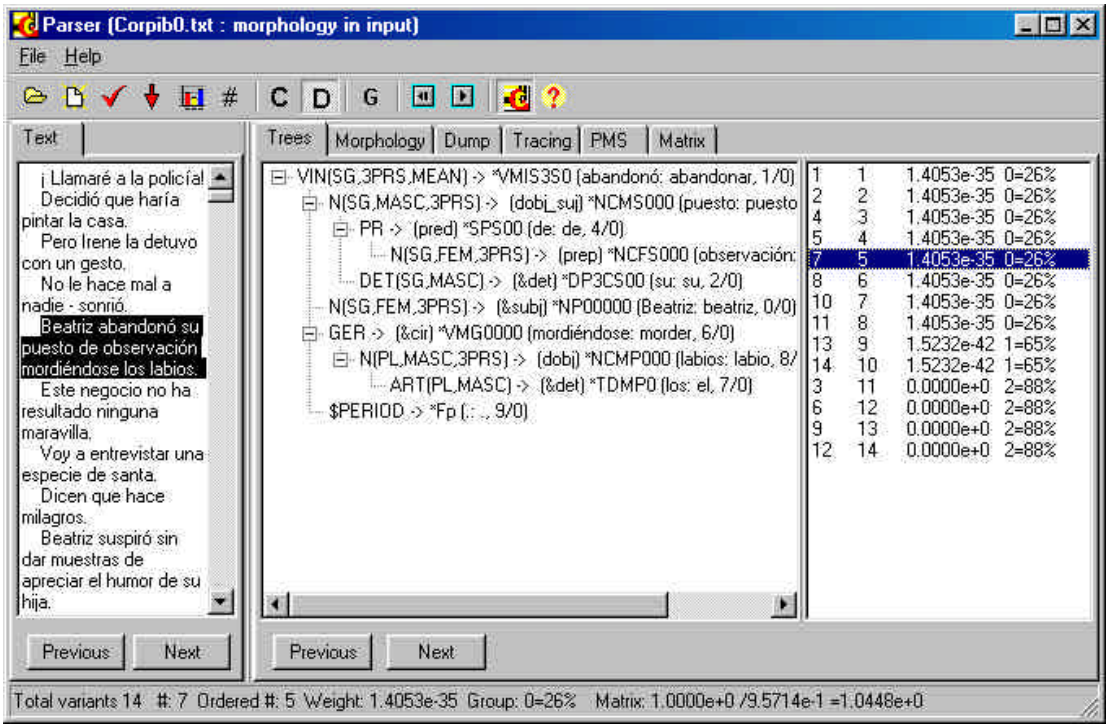
1) Una oración donde los pesos de las combinaciones determinan la variante correcta, la oración es *Voy a entrevistar una especie de santa*. En la parte izquierda la primera columna indica la posición de la variante en la salida del analizador básico, la segunda columna la posición de la variante debido a los pesos de las combinaciones, la tercera columna el peso de la variante y la cuarta columna el grupo de variantes con el mismo peso y su porcentaje de colocación.



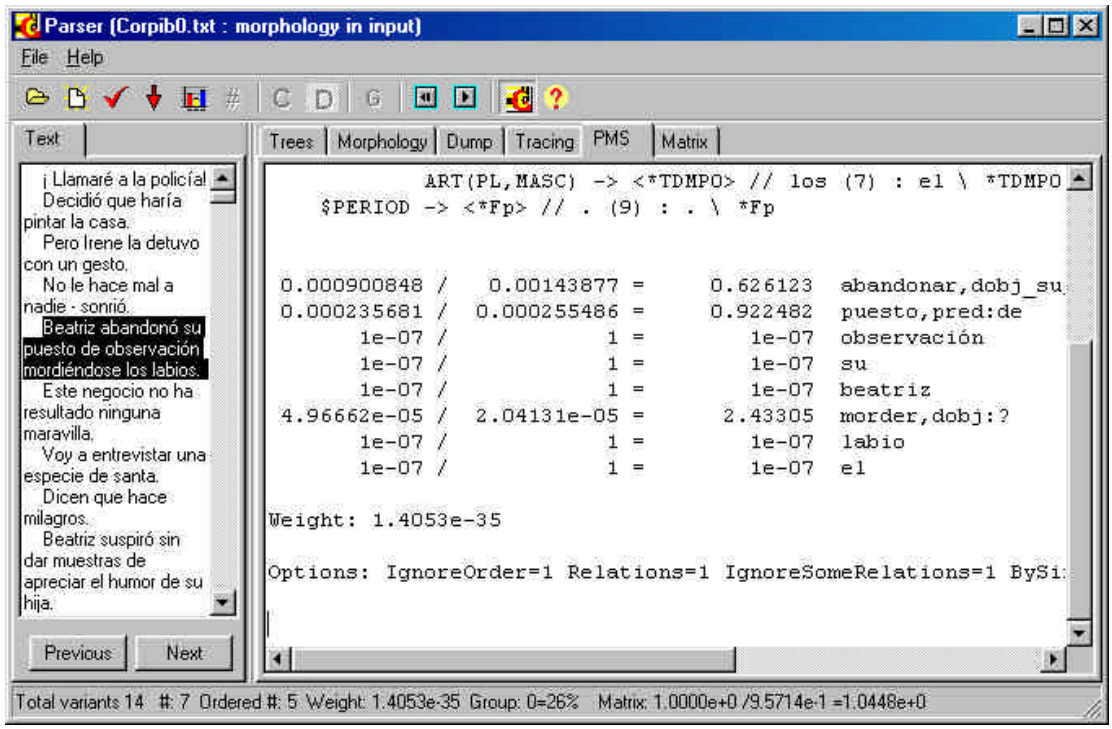
Los valores de las combinaciones que logran este resultado se presentan enseguida, donde es de notar el peso de la combinación "ir, obj:*a*"



2) El ejemplo a continuación, para la frase *Beatriz abandonó su puesto de observación mordiéndose los labios*, muestra un caso de falla de posición por valores muy bajos de las combinaciones.



La variante correcta tiene la posición 5 de 14 variantes. Las combinaciones y sus pesos se muestran a continuación donde se observa el peso muy bajo para la combinación *abandonar algo*, por la falta de datos en las oraciones analizadas del corpus. En este ejemplo, la diferencia entre la variante 1 y la 5 es la relación de la frase *mordiéndose los labios* respecto al verbo. En la primera variante como un modificador del verbo y en la segunda variante como circunstancial de la oración: *abandonó su puesto de observación.*


Parser (Corpib0.txt : morphology in input)
File Help
Text
¡ Llamaré a la policía!
Decidió que haría pintar la casa.
Pero Irene la detuvo con un gesto.
No le hace mal a nadie - sonrió.
Beatriz abandonó su puesto de observación mordiéndose los labios.
Este negocio no ha resultado ninguna maravilla.
Voy a entrevistar una especie de santa.
Dicen que hace milagros.
Beatriz suspiró sin dar muestras de apreciar el humor de su hija.
Previous
Next
Trees
Morphology
Dump
Tracing
PMS
Matrix
ART(PL,MASC) -> <*TDMPO> // los (7) : el \ *TDMPO
$PERIOD -> <*Fp> // . (9) : . \ *Fp
0.000900848 / 0.00143877 = 0.626123 abandonar,dobj_su
0.000235681 / 0.000255486 = 0.922482 puesto,pred:de
1e-07 / 1 = 1e-07 observación
1e-07 / 1 = 1e-07 su
1e-07 / 1 = 1e-07 beatriz
4.96662e-05 / 2.04131e-05 = 2.43305 morder,dobj:?
1e-07 / 1 = 1e-07 labio
1e-07 / 1 = 1e-07 el
Weight: 1.4053e-35
Options: IgnoreOrder=1 Relations=1 IgnoreSomeRelations=1 BySi
Total variants 14 #: 7 Ordered #: 5 Weight: 1.4053e-35 Group: 0=26% Matrix: 1.0000e+0 /9.5714e-1 =1.0448e+0

4.11 RESULTADOS DE LA APLICACIÓN DE LOS PESOS DE COMBINACIONES EN EL ANALIZADOR BÁSICO

Compilamos las combinaciones y sus pesos en una base de datos para ser utilizada por el analizador básico ya descrito. Con estos valores el analizador básico calcula los pesos de las diversas variantes que generó. Las variantes se clasifican de acuerdo a los pesos obtenidos para las variantes. Esta clasificación promueve el grupo de variantes con mayor posibilidad de ser el correcto a la posición tope de la lista de variantes clasificadas.

Para realizar la prueba de efectividad del método, tomamos un conjunto de oraciones del corpus LEXESP, con menor número de variantes. Este conjunto se presenta en el Anexo Conjunto de prueba. El método simple que consideramos es el siguiente:

1. Determinamos la variante correcta de entre todas las variantes generadas.
2. El analizador básico construido, clasifica las variantes mediante los pesos de las combinaciones.
3. Anotamos la posición de la variante correcta en la clasificación anterior y calculamos el rango medio de esa posición respecto al total de variantes.

Cabe hacer notar que en el analizador básico la posición de salida no tiene ninguna relación con su posibilidad de ser la correcta sino con el número de marca morfológica seleccionada para cada palabra y con la longitud y orden alfabético de las reglas empleadas. En la salida obtenida al aplicar los pesos de las combinaciones, la posición se debe únicamente a nuestro método.

En la tabla siguiente se observan los pesos obtenidos para 53 de las 100 oraciones del conjunto de prueba.

Oración	**# posición variante correcta**	**Rango medio**	**#Total de variantes**
1	2	50%	3
2	1	0%	14
3	4	15%	20
4	5	9%	44
5	5	26%	14
6	1	0%	2
7	1	0%	169
8	3	100%	3
9	669	40%	1660
10	25	5%	480
11	73	61%	118
12	---	---	Mal analizada
13	441	13%	3144
14	555	59%	936
15	3	4%	48
16	2	33%	4
17	---	---	Mal analizada
18	1	0%	42
19	1	0%	10
20	1	0%	288
21	1	0%	6
22	28	31%	88
23	25	14%	170
24	17	41%	40
25	---	---	160200
26	1	0%	12
27	1	0%	6

Resultados de la aplicación de los pesos de combinaciones en el analizador básico

Oración	# posición variante correcta	Rango medio	#Total de variantes
28	1	0%	15
29	---	---	Mal analizada
30	---	---	Mal analizada
31	4	42%	8
32	---	---	Demasiadas 6360
33	---	---	Mal analizada
34	4	60%	6
35	---	---	Una variante
36	---	---	Mal analizada
37	1	0%	18
38	1	0%	11
39	---	---	Una variante
40	5	12%	32
41	---	---	Mal analizada
42	---	---	Mal analizada
43	---	---	Mal analizada
44	2	100%	2
45	1	0%	26
46	5	17%	24
47	---	---	Mal analizada
48	5	57%	8
49	1	0%	16
50	---	---	Mal analizada
51	1	0%	4
52	361	82%	440
53	19	56%	33

De los valores obtenidos, se observa que con nuestro método, se logra como rango medio de colocación el 25%.

CONCLUSIONES

Motivación

El problema del análisis sintáctico y la desambiguación de las estructuras sintácticas generadas es un elemento importante en el análisis lingüístico de textos por computadora. Los analizadores sintácticos que se han construido con una base puramente gramatical generan tal cantidad de variantes que su empleo resulta casi inútil.

Para eliminar esa gran cantidad de variantes incorrectas se han adicionado distintos métodos, entre ellos las restricciones en los formalismos gramaticales, una noción muy importante de la gramática universal. Con el mismo fin, se han incluido otros métodos en los analizadores, principalmente métodos estadísticos para obtener las probabilidades de concurrencias de palabras o categorías gramaticales. Sin embargo, para resolver la desambiguación de estructura sintáctica se requiere proveer a la máquina, con el conocimiento lingüístico que los hablantes nativos poseen, absorbido en los años iniciales del aprendizaje del primer lenguaje. Este conocimiento lingüístico está asociado con fuentes de conocimiento léxico, sintáctico y semántico.

En esta tesis proponemos un nuevo modelo de análisis sintáctico y desambiguación para el español. El analizador sintáctico incluye un esquema de diferentes fuentes de conocimiento, cada una como un grado de libertad en un dominio específico. La desambiguación estructural se basa en la contribución mayoritaria de las evaluaciones cuantitativas de cada una de las variantes, todas en un formato compatible.

El enfoque que tomamos para resolver este problema de ambigüedad estructural considera los siguientes aspectos: introducir fuentes de conocimiento léxico, sintáctico y semántico, representar este conocimiento en diccionarios cuya compilación sea automática en su mayor parte, desarrollar algoritmos muy simples y eficientes para todas las tareas necesarias, y el uso recursivo de las herramientas desarrolladas.

Contribuciones

Las mayores contribuciones de esta tesis son las siguientes:

Comparar los enfoques existentes, para caracterizar las valencias sintácticas del español.

Comparamos los métodos de descripción de estructuras sintácticas, desarrollados en los enfoques de constituyentes y de dependencias. En el enfoque de constituyentes o gramáticas generativas incluimos la Teoría de Rección y Ligamento (GB), la Gramática de Estructura de Frase Generalizada (GPSG), la Gramática Léxica Funcional (LFG), la Gramática Categorial, y la Gramática de Estructura de Frase

dirigida por el núcleo-*h* (HPSG). En el enfoque de dependencias, consideramos la Gramática de Unificación de Dependencias (DUG) y la Teoría Texto ⇔ Significado (MTT).

Proponer un enfoque generalizado para la descripción de valencias. La descripción une los méritos de los enfoques existentes.

Basándonos en la comparación arriba mencionada, y en algunas características del español, como el orden más libre de palabras, la dependencia del objeto directo en la animidad, repetición limitada de los objetos, el complemento beneficiario y su duplicación, mostramos la necesidad de caracterizar las valencias de los verbos, de los adjetivos y de algunos sustantivos del español. El enfoque generalizado, propuesto para describir de una manera más natural y adecuada las valencias, incluye características consideradas en las gramáticas generativas y en las gramáticas de dependencias, aunque con mayor énfasis en la descripción sintáctica de la Meaning ⇔ Text Theory.

Elaborar un esquema de diferentes fuentes de conocimiento para la generación de variantes sintácticas en el proceso de análisis sintáctico.

El esquema de análisis sintáctico que proponemos considera la inclusión de tres fuentes de conocimiento: léxica, sintáctica y semántica. Sólo con la participación de estos conocimientos es posible diferenciar las variantes sintácticas correctas de entre las múltiples variantes sintácticas generadas en el proceso de análisis sintáctico, es decir, eliminar el mayor número posible de variantes erróneas de las diversas fuentes.

Las fuentes de conocimiento son:

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

El método principal de análisis sintáctico se basa en un diccionario tipo Patrones de manejo, que es la descripción sintáctica en la MTT. Proponemos una nueva estructura formal de los patrones de manejo. Además de modernizar su formato, nos basamos en los sistemas de análisis sintáctico que dan mayor importancia a los diccionarios para representar cada característica con el sistema de pares de atributo valor. Incluimos también información de evaluación estadística.

La compilación de este tipo de diccionarios, hasta ahora solamente ha sido posible manualmente. Para eliminar esta desventaja elaboramos un algoritmo para compilar automáticamente las combinaciones de subcategorización que forman las valencias.

- *Reglas ponderadas*, que reflejan conocimiento sintáctico.

El método de reglas ponderadas se basa en una gramática generativa extendida con rasgos de concordancia (género, número, persona). Para hacer del método la herramienta básica de análisis sintáctico introdujimos varios elementos. La inclusión

del elemento rector en cada regla nos permite hacer una transformación de estructura de constituyentes a estructura de dependencias. La inclusión de relaciones sintácticas, además de establecer la dirección de las dependencias, permite diferenciar entre valencias y algunos complementos circunstanciales.

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

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

El empleo de la red semántica para la desambiguación sintáctica tiene la finalidad de incorporar la componente semántica faltante en las otras dos fuentes. La estructura sintáctica en este modelo se toma de la salida producida con las reglas ponderadas. Algunas de las gramáticas más actuales, derivadas de las gramáticas generativas precisamente incorporan restricciones semánticas, como la HPSG que las considera en la entrada de cada lexema en el diccionario, lo cual implica una labor manual, intensiva en extremo.

En nuestro modelo, esas restricciones semánticas se buscan en la red y se definen a través de la proximidad semántica, que involucra la distancia menor entre pares de palabras y su valor asignado. La evaluación de la proximidad no nada más está relacionada con estos valores obtenidos de la red misma, sino que considera además el tipo sintáctico de la relación.

Seleccionar y realizar un analizador básico que permite utilizar las tres fuentes de conocimiento para mejorar la desambiguación de las variantes sintácticas.

Seleccionamos un algoritmo ascendente de análisis sintáctico para CFG, con tabla para guardar resultados intermedios, y realizamos un analizador sintáctico basado en la gramática de reglas ponderadas. Este analizador básico incluye una característica importante: la transformación de sus estructuras de salida, de estructuras de constituyentes a estructuras de dependencias, mediante la cual se obtiene un mismo formato para las variantes de salida de las tres fuentes.

De cada fuente de conocimiento, la salida resultante es un grupo de diversas variantes que no están ordenadas. Para ordenarlas en cuanto a su posibilidad de ser la variante correcta, asignamos un peso a cada variante, así que cada una de las fuentes contribuye con variantes distinguidas con pesos específicos. La asignación de pesos a las variantes se da de acuerdo a los pesos de combinaciones que forman las variantes,

y a características específicas de los métodos que las producen. Para realizar este ordenamiento, proponemos dos métodos de votación, una forma simple de mayor contribución, y la posibilidad de incluir un módulo de evaluación múltiple. De esta forma, las variantes que se encuentran más al inicio de la salida del módulo de votación son las variantes con mayor posibilidad de ser las correctas. Sin la transformación a un formato compatible no sería posible determinar las variantes sobresalientes porque sus valores no serían comparables.

Elaborar un algoritmo y un sistema para acumular las estadísticas de las variantes sintácticas. Este sistema utiliza el analizador básico como una herramienta.

El desarrollo del sistema de compilación del diccionario de patrones de manejo y la generación de variantes mediante reglas ponderadas comparten la misma herramienta, es decir, el método de reglas ponderadas se emplea recursivamente. Nos basamos en la intuición y en experimentos preliminares para hacer más eficientes los programas de compilación.

El algoritmo se basa en métodos lingüísticos y estadísticos, el método lingüístico es un formalismo generativo. Los métodos estadísticos se emplean para la selección de las combinaciones que forman las variantes de estructuras, con la finalidad de obtener los complementos de palabras específicas (verbos, adjetivos y algunos sustantivos). La base del algoritmo son las fórmulas desarrolladas para la obtención de los pesos estadísticos de las diferentes variantes de análisis sintáctico de una frase. Los pesos de las variantes se basan a su vez en los pesos de las combinaciones de subcategorización que aparecen en los árboles sintácticos.

Para obtener los pesos iniciales de las variantes empleamos información de patrones elaborados manualmente. Basándonos en algunos pesos calculamos los pesos de las variantes, y con estos valores asignamos nuevos pesos a las combinaciones. Estos nuevos pesos de combinaciones permiten obtener a su vez nuevos pesos para cada una de las variantes de la frase, por lo que un proceso iterativo sobre un corpus de textos da por resultado la obtención de la acumulación de los pesos de las combinaciones de subcategorización que forman los patrones de manejo.

El algoritmo permite reducir significativamente el problema de compilación del diccionario de patrones de manejo, a pesar de ser extremadamente sencillo y solamente débil estadísticamente.

Emplear las estadísticas obtenidas como el conocimiento modernizado sobre los patrones de manejo.

Con las estadísticas obtenidas podemos conocer las probabilidades: de diferentes opciones de llenado de diferentes valencias, del uso de diferentes opciones de la misma valencia, y de compatibilidad de varias combinaciones de opciones específicas para diferentes valencias.

Así que mediante el algoritmo de compilación obtenemos mayor evaluación estadística para eliminar cierta ambigüedad en el análisis sintáctico y para favorecer determinadas realizaciones que aparecen con mayor frecuencia en corpus de textos, lo cual no ha sido considerado en compilaciones de este tipo de diccionarios.

Comparar las cualidades del analizador básico antes de emplear estas estadísticas y después de introducirlas.

El analizador básico se construyó con la característica adicional de incorporar los resultados directos de la compilación de patrones. Con esta información es posible clasificar las variantes generadas de tal forma que se promueven hacia el tope de la clasificación a las variantes con mayor posibilidad de ser las correctas.

Entonces, para medir la efectividad de los pesos de las combinaciones de subcategorización obtenidas, aplicamos nuestros resultados al analizador sintáctico básico que clasifica las variantes conforme a los pesos incorporados de las combinaciones. El resultado es que logramos un rango medio de 25% de colocación de la variante correcta en esta clasificación.

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

Rumbos de investigación posteriores

La adquisición de conocimiento lingüístico por computadora y la desambiguación estructural son tareas difíciles, aunque creemos que las investigaciones reportadas en esta tesis representan cierto progreso, nos es claro que se requiere mayor investigación para resolver estos problemas. Algunos elementos que quedan para investigaciones posteriores son los siguientes:

En el desarrollo del algoritmo de compilación de combinaciones de valencias sintácticas, consideramos un método estadístico basado en hipótesis de independencia, principalmente la introducción de las combinaciones de las estructuras incorrectas de forma independiente a la estructura correcta de la frase, y la independencia entre combinaciones de una frase. La obtención de datos cuantitativos útiles sobre esas dependencias debe realizarse para investigar su influencia en los resultados obtenidos.

En esta tesis proponemos un modelo basado en tres tipos de conocimiento, cada uno con un módulo individual. En lugar de tres módulos separados podría proponerse un modelo unificado. El problema de cómo definirlo es un trabajo a futuro.

La cantidad de datos disponibles es menor que la necesaria, está pendiente de solución el problema de cómo reunir la información requerida, para nuestro caso: frases con todos los objetos de los lexemas. Por lo que consideramos que deben investigarse métodos de extracción particularizada.

GLOSARIO

Constituyente: elemento lingüístico que forma parte de una construcción superior donde las oraciones se analizan mediante un proceso de segmentación y clasificación. Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso para cada parte dividiéndola en subconstituyentes, y así sucesivamente hasta que las partes sean las partes de la palabra indivisibles dentro de la gramática (morfemas).

Clítico: elemento átono fonológicamente dependiente de otro dotado de acento, ejemplos: los pronombres *me*, *te le*, etc.

Concordancia: en muchos lenguajes, las formas de ciertos elementos pueden variar para indicar propiedades de persona, número, género, etc. Estas variaciones a menudo se describen por afijos. Algunas relaciones gramaticales entre pares de elementos requieren el acuerdo entre estas propiedades.

Coordinación: se refiere a la unión de dos palabras o frases de equivalente condición sintáctica.

Desambiguación: eliminación de ambigüedades.

Descriptivo (método): estudio de la estructura o funcionamiento de una lengua o dialecto sin atender a su evolución, es decir, sin considerar los fenómenos que ocurren a lo largo del tiempo, evaluando los datos objetivamente definibles o mensurables.

Ditransitividad o doble transitividad. El esquema típico para la doble transitividad en español es verbo seguido de objeto directo, seguido de objeto indirecto.

Especificadores: término que cubre sujetos de oraciones, determinantes de grupos nominales y cierta clase de constituyentes que no son núcleos ni complementos de los núcleos.

Extraposición: en este fenómeno lingüístico se mueven ciertos complementos del tipo nominal a la posición final de la oración y se sustituyen con un pronombre vacío.

Gramaticalidad: cualidad de una secuencia oracional, por la que se ajusta a las reglas de la gramática.

Lematización: reducción de las formas flexivas de los lexemas que aparecen en un texto, a su respectivo lema o forma de cita convencional. Por ejemplo: las formas *amo*, *amas*, *aman*, en su lema *amar*.

Lexema: unidad léxica abstracta que no puede descomponerse en otras menores aunque si 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*.

Lexicalismo (lexicismo): a menudo se refiere a la teoría de que la estructura interna de las palabras es independiente de cómo se juntan las palabras para hacer oraciones y de que las plabras son los átomos de las combinaciones sintácticas. Está relacionado 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.

Mapear: es una forma de asociar objetos únicos a cada punto de un conjunto dado.

Morfema: palabra de la terminología gramatical moderna con que se designan los elementos lingüísticos que se incorporan a las palabras con significado fijo y forma variable. Morfema puede ser una palabra, prefijo, infijo, sufijo, desinencia, etc.

No terminales: son variables sintácticas que denotan conjuntos de frases o cadenas de palabras. Estos conjuntos ayudan a definir el lenguaje generado por la gramática imponiéndole una estructura jerárquica. SE corresponden con las categorías gramaticales.

Prescriptivo (método): en oposición a descriptivo, método que propone y sanciona ciertas normas lingüísticas consideradas canónicas al tiempo que condena los usos desviados y las innovaciones procedentes de cualquier otro modelo.

Recursividad: método matemático para definir funciones, que consiste en partir de una base e ir construyendo los componentes de la función haciendo referencia a la definición de la función misma, en una especie de "círculo vicioso controlado". Familia: recursión, recursivo.

Rema: lo que se dice del tema.

Subcategorización: clasificación rigurosa, sistemática y jerárquica, según rasgos de las unidades léxicas de la lengua, para describir cuántos y de que tipo son los elementos con los que combina para hacer oraciones completas. Cuando se dice que subcategoriza determinada categoría gramatical, significa que combina con ella.

Subsumir: Incluir algo como componente en una síntesis o clasificación más abarcadora.

Tema: aquello de lo que se habla en la oración (sujeto sicológico).

Terminales son los símbolos básicos con que se forman las frases del lenguaje. Coinciden más o menos con las palabras de una lengua, y se agrupan en el diccionario.

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*.

Unificación: 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.

Verbos de ascensión: como el verbo *seems* (parecer) que introduce otro verbo como predicadoy donde se considera que cada verbo tiene un sujeto, incluso el infinitivo. Se denomina sujeto de ascensión (*subject raising*, en inglés) si es transparente en cuanto a que el sujeto también es sujeto del verbo que introduce. Se denomina objeto de ascensión, (*object raising*, en inglés) si el objeto es el sujeto del verbo que introduce.

Verbos de control o verbos *equi*. En estos verbos que introducen otros grupos verbales, el sujeto no es transparente. El controlador y el controlado son ambos temáticos.

Verbo finito (o en forma finita): es un verbo que tiene marcas de tiempo.

VOCABULARIO BILINGÜE DE TÉRMINOS (INGLÉS – ESPAÑOL)

actuante	*actant*
ambigüedad	*ambiguity*
análisis	*analysis*
analizador sintáctico ascendente	*botom-up parsing*
analizador sintáctico descendente	*top-down parsing*
analizador sintáctico	*syntactic analyzer*
analizador	*analyzer*
árbol de constituyentes	*constituent tree*
árbol de dependencias	*dependency tree*
ascensión del objeto	*object raising*
ascensión del sujeto	*subject raising*
cadena	*string*
caso gramatical	*grammatic case*
CFG	context-free grammars+
comprensión de lenguaje natural	*natural language understanding*
constituyente	*constituent*
dependencia	*dependency*
desambiguación	*disambiguation*
estructura de frase	*phrase structure*
estructura profunda	*deep structure*
estructura sintáctica	*syntactic structure*
fonología	*phonology*
forma de la palabra	*wordform*
frase	*phrase*
generación	*generation*
Gramática de Estructura de Frase dirigida por el Núcleo	*Head-driven Phrase Structure Grammar*
Gramática Generalizada de Estructura de Frase	*Generalized Phrase Structure Grammar*
gramáticas generativas	*generative grammars*
gramáticas libres de contexto	*context-free grammars*
gramáticas transformacionales	*transformational grammars*
homonimia	*homonymy*
homónimo	*homonym*

HPSG	ver *Head-driven Phrase Structure Grammar*
lexema	*lexeme*
lexicografía	*lexicography*
lingüística sociológica	*sociolinguistcs*
marco de subcategorización	*subcategorization frame*
morfología	*morphology*
morfosintáctico	*morphosyntactic*
no terminal	*nonterminal*
núcleo	*head*
parser (analizador sintáctico)	*parser*
partes del habla (de la oración)	*part of speech*
patrón de manejo o rección	*government pattern*
polisemántico	*polysemic*
polisemia	*polysemy*
predicado sintáctico	*syntactic predicate*
red semántica	*semantic network*
reescribir	*rewritting*
rema	comment
restricción	*constraint*
semántica	*semantics*
sicolingüística	*psycholinguistics*
signo lingüístico	*linguistic sign*
sinonimia	*synonymy*
sinónimo	*synonym*
sintáctica	*syntactics*
sintaxis	*syntax*
síntesis	*synthesis*
tema	topic
Teoría Significado ⇔ Texto	*Meaning ⇔ Text theory*
unificación	*unification*
valencia	*valency*

LISTA DE TÉRMINOS

LISTA DE PUBLICACIONES DE LA TESISTA SOBRE EL TEMA DE TESIS

Revistas indexadas por SCI

1) I. A. Bolshakov , S. N. Galicia-Haro. *Algoritmo para corregir ciertos errores ortográficos en español relacionados al acento.* J. **Computación y Sistemas** (incluida en el Indice del CONACyT, **indexada por SCI**). No.2 1997, México.

2) A. Bolshakov, A.F. Gelbukh, S.N. Galicia Haro. *Electronic Dictionaries: For both Humans and Computers* (also published in Russian). J. **International Forum on Information and Documentation** FID 519 (**indexada por SCI**), ISSN 0304-9701, N 3, 1999.

Otras revistas

3) S. N. Galicia Haro *Corrección de estilo en lenguajes Naturales.* Revista **Soluciones Avanzadas**. Abril de 1995. México.

4) S. N. Galicia Haro, I. Bolshakov, A. Gelbukh. *Diccionario de patrones de manejo sintáctico para análisis de textos en español.* J ***Procesamiento de Lenguaje Natural*** N 23, September 1998 (indexada por INIST-CNRS, Institut de L'information Scientifique et Technique, Centre National de la Recherche Scientifique), ISSN 1135-5948.

Capítulos en libros de memorias de *Springer*

5) Igor A. Bolshakov, Alexander F. Gelbukh, Sofia N. Galicia-Haro. *Electronic Dictionaries: for both Humans and Computers*. Accepted to Proc. Workshop on Text, Speech and Dialog (TSD-99), Pilsen, Czech Republic, September 13-17, 1999, Lecture Notes in Artificial Intelligence N 1692, **Springer-Verlag**, ISBN 3-540-66494-7.

6) S. N. Galicia-Haro, I. A. Bolshakov, A. F. Gelbukh. *A Simple Spanish Part of Speech Tagger for Detection and Correction of Accentuation Errors*. Accepted to Proc. Workshop on Text, Speech and Dialog (TSD-99), Pilsen, Czech Republic, September 13-17, 1999. Lecture Notes in Artificial Intelligence N 1692, **Springer-Verlag**, ISBN 3-540-66494-7.

Capítulos en libros de *Selected Papers*

7) I.A. Bolshakov , S. N. Galicia-Haro. *Detección y corrección de ciertos errores ortográficos en español relacionados al acento*. In: A. de Albornoz Bueno, A. Guzmán Arenas *et al.* (Eds.) *Selected Works 1997-1998*, ISBN 970-18-3427-5. (Una versión revisada del trabajo N 1)

8) S. N. Galicia Haro, I. Bolshakov, A. Gelbukh. *Descripción y aplicación de un diccionario de patrones de manejo sintáctico para el español*. In: A. de Albornoz Bueno, A. Guzmán Arenas *et al.* (Eds.) *Selected Works 1997-1998.* (Una versión revisada del trabajo N 4)

9) S.N. Galicia-Haro, A.F. Gelbukh, I.Bolshakov. *The Structure of Spanish Linguistic Processor*. In: A. de Albornoz Bueno, A. Guzmán Arenas *et al.* (Eds.) *Selected Works 1997-1998.* (Una versión revisada del trabajo N 16)

10) S. N. Galicia-Haro. I.A. Bolshakov, A.F. Gelbukh. *Patrones de manejo sintáctico para verbos comunes del español*. In: A. de Albornoz Bueno, A. Guzmán Arenas *et al.* (Eds.) *Selected Works 1997-1998.* (Una versión revisada del trabajo N 17)

11) A. Bolshakov, A.F. Gelbukh, S.N. Galicia Haro. *Towards the Idea of the Universal Dictionary*. In: A. Guzmán Arenas, F. Menchaca García (Eds.) *Selected Papers. Computing Research 1999*. (Una versión revisada del trabajo N 2)

12) Igor A. Bolshakov, Alexander F. Gelbukh, Sofia N. Galicia-Haro. *How to Improve the Electronic Dictionaries*. In: A. Guzmán Arenas, F. Menchaca García (Eds.) *Selected Papers. Computing Research 1999*. (Una versión revisada del trabajo N 5)

13) S. N. Galicia-Haro, I. A. Bolshakov, A. F. Gelbukh. *Correcting Some Spanish Accentuation Errors by means of Context Noun Group Detection*. In: A. Guzmán Arenas, F. Menchaca García (Eds.) *Selected Papers. Computing Research 1999*. (Una versión revisada del trabajo N 6)

Congresos internacionales

14) S. N. Galicia-Haro. *Un diccionario combinatorio para el Español. Primeros pasos.* Conferencia DIALOG-97 sobre Lingüística Computacional. Moscú, Rusia, Junio de 1997.

15) S. N. Galicia-Haro, A.F. Gelbukh. *A combinatorial dictionary of Spanish: the first steps*. Dialogue-97, pp. 68 - 70, Annual International Conf. on Applied Linguistics, Moscow, RIAI, 1997.

16) S.N. Galicia-Haro, A.F. Gelbukh, I.Bolshakov. *The "Understanding" of Spanish Texts: CIC Language Understanding Project*. Proc. International Workshop on Spanish Language Processing, Santa Fe, New Mexico, USA, July 21 - 22, 1997.

17) S. N. Galicia-Haro. I.A. Bolshakov, A.F. Gelbukh. *Patrones de manejo sintáctico para verbos comunes del español*. CIC'97, Nuevas Aplicaciones e Innovaciones Tecnológicas en Computación, Simposium internacional de computación, 12-14 de noviembre, pp. 367 - 371, 1997, CIC, IPN, México D.F.

18) A. Gelbukh, I. Bolshakov, S. N. Galicia Haro. *Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts* // International Conference on Automatic Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA, USA, June 11 - 13, pp. 261 - 267, 1998.

19) I. Bolshakov, A. Gelbukh, S. N. Galicia Haro. *Simulation in linguistics: assessing and tuning text analysis methods with quasi-text generators* // Proc. of the Annual International Conf. on Applied Linguistics "Dialogue-98", November 1998, Moscow, Russia, pp. 768 - 775.

20) A. Gelbukh, I. Bolshakov, S. Galicia-Haro. *Statistics of parsing errors can help sintactic disambiguation*. CIC-98 - Simposium Internacional de computación, November 11 - 13, 1998, Mexico D.F., pp. 405 – 515.

21) A.F. Gelbukh, Sofia N. Galicia-Haro. *An extended subcategorization frames dictionary* (abstract). Program of 30th Annual Conference of Canadian. Association of Applied Linguistic in conjunction with Congress of the Social Scienties and Humanities, Sherbrooke, Canada, June 3-5, 1999

22) Igor Bolshakov, Alexander Gelbukh, and Sofia Galicia-Haro. *A Simple Method to Detect and Correct Spanish Accentuation Typos*. Proc. PACLING-99, Pacific Association for Computational Linguistics, University of Waterloo, Waterloo, Ontario, Canada, August 25-28, 1999, ISBN 0-9685753-0-7, pp. 104-113. http://www.lpaig.uwaterloo.ca/ ~b2hui/pacling/presenters.html

23) I. Bolshakov, P. Cassidy, A.Gelbukh, G. Sidorov, S. Galicia-Haro. *'Non-adult' semantic field*. Accepted to Proc. 3rd Tbilisi Symposium on Language, Logic,

and Computation. Batumi, Georgia, September 12–16, 1999 (http:// www.illc.uva.nl/ Batumi)

24) S. N. Galicia-Haro, A.F. Gelbukh, I.A. Bolshakov. *Advanced Subcategorization Frames for Languages with Relaxed Word Order Constraints (on Spanish Examples).* November 22-24 International Natural Language Processing VEXTAL. 1999, pp. 101- 110.

25) Sofía N. Galicia-Haro, I. A. Bolshakov, and A. F. Gelbukh. *Aplicación del formalismo Meaning ⇔ Text Theory al análisis de textos en español*. CIC-99, Simposium Internacional de Computación, November 15 - 19, 1999, CIC, IPN, Mexico D.F., pp. 342-351.

26) Sofía N. Galicia-Haro, I. A. Bolshakov, and A. F. Gelbukh. *Un modelo de descripción de la estructura de las valencias d everbos españoles para el análisis automático de textos.* Avances en Inteligencia Artificial. Mexican International Conference on Artificial Intelligence. III Taller de Inteligencia Artificial. V Taller Iberoamericano de Reconocimiento de Patrones MICAI/TAINA/TIARP 2000.

27) G. O. Sidorov, I. A. Bolshakov, P. Cassidy, S. Galicia-Haro, A. F. Gelbukh. *A Comparative analysis of the semantic field "non-adult" in Russian, English, and Spanish*. Accepted to Proc. Annual International Conf. on Applied Linguistics Dialogue-2000, June, 2000, Moscow, Russia.

Congresos nacionales

28) S. N. Galicia-Haro. *Ayuda de corrección de estilo para textos en español, traducidos del inglés.* Congreso Nacional de Informática. Ponencia técnica. Junio 1994.

29) S.N. Galicia-Haro, A.F.Gelbukh, I.A. Bolshakov. *Syntactical government pattern dictionaries in language teaching and automatic text processing*. Proc. COPEI-97, Congreso Nacional sobre Educación en Ingeniería y Desarrollo Sustentable, November 26-29, 1997, Morelia, México.

30) A. Gelbukh, S. Galicia-Haro, I.Bolshakov. *Three dictionary-based techniques of disambiguation*. TAINA-98, Workshop on Artificial Intelligent, Mexico D.F., pp. 78 – 89, ISBN 970-18-2057-6.

31) S. N. Galicia-Haro, A. F. Gelbukh, e I. A. Bolshakov. *Un método de descripción de conocimiento lingüístico y su aplicación al análisis sintáctico*

del español. En Memorias del Segundo Encuentro de Computación 1999, Septiembre 1999. http://titan.udlap.mx/~enc99/ listaLogica.html

32) S. N. Galicia-Haro. *Desambiguación sintáctica para el español, basada en patrones de manejo*. TAINA-99, Taller de Inteligencia Artificial, Mexico D.F., pp. 164 – 178, ISBN 970-1835549.

Informes Técnicos

33) A. Gelbukh, I. Bolshakov, S. N. Galicia Haro. *Syntactic Disambiguation by Learning Weighted Government Patterns from a Large Corpus*. Technical report. CIC, IPN, 1998.

34) I. Bolshakov, A. Gelbukh, S. Galicia Haro, M. Orozco Guzmán. *Government patterns of 670 Spanish verbs*. Technical report. CIC, IPN, 1998.

35) A.F. Gelbukh, I.A. Bolshakov, Sofia N. Galicia-Haro, M.A. Alexandrov, P.P. Makagonov, P.J. Cassidy. *Dictionaries for text processing and language teaching: use and compilation*. Technical report. CIC, IPN, ISBN 970-18-3322-8, 1999.

Conferencias impartidas

36) *Three Dictionary-Based Techniques of Disambiguation* (con A. Gelbukh). 9th Conference on Engineering, September 29 to October 3, 1997, Instituto Tecnológico Autónomo de México (ITAM), Mexico City.

37) *Las valencias sintácticas y semánticas en el análisis automático del texto en español. 6ª* Conferencia, 21 de enero del 2000, Seminario de Investigación CIC-IPN, México, D. F.

REFERENCIAS

[Abney, 91] Abney, S. P. *Parsing by chunks*. In R. C. Berwick, S. P. Abney, and C. Tenny, editors, Principle-Based Parsing: Computation and Psycho-linguistics, pages 257--278. Kluwer, Dordrecht, 1991.

[Agirre & Rigau, 96] Agirre, E. and Rigau, G. *Word Sense Disambiguation using Conceptual Density*. In Proceedings of the 16th International Conference on Computational Linguistics (COLING96). Copenhagen, Denmark, 1996. http://xxx.lanl.gov/ps/cmp-lg/9606007

[Aho *et al*, 86] Aho, A. V., R. Sethi and J. D. Ullman. *Compilers. Principles, Techniques and Tools*. Addison Wesley Publishing Company, 1986.

[Ajdukiewicz, 35] Ajdukiewicz, K. *Die syntaktische Konnexitùat*. Studia Philosophica 1, 1--27, 1935.

[Alarcos, 84] Alarcos-Llorach, E. *Gramática Estructural*. Editorial Gredos. Madrid, 1984.

[Allen, 95] Allen, J. F. *Natural Language Understanding*. Benjamin Cummings, 1995.

[Alonso, 60] Alonso Pedraz, M. *Diccionario Ideoconstructivo*. En Ciencia del Lenguaje y Arte del Estilo. Editorial Aguilar. Madrid, España, 1960.

[Alsina, 93] Alsina, A. *Predicate Composition: A Theory of Syntactic Function Alternations*. PhD thesis, Stanford, 1993.

[Anttila, 95] Anttila, A. *How to recognise Subjects in English*. In Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (eds.) Constraint grammar: A Language-Independent System for Parsing Unrestricted text. Mouton de Gruyter, 1995.

[Apresyan *et al*, 73] Apresyan, Yu. D., I. A. Mel'cuk and A. K. Zolkovsky. *Materials for an explanatory combinatory dictionary of modern Russian*. In Trends in Soviet Theoretical Linguistics. Edited by F. Kiefer. Foundations of Language Supplementary Series., vol. 18. Reidel, Dordrecht, 1973.

[Argamon *et al*, 98] Shlomo Argamon, Ido Dagan and Yuval Krymolowski. *A Memory-Based Approach to Learning Shallow Natural Language Patterns*. In Proceedings Intern. Conference COLING-ACL'98. August 10-14 Quebec, Canada. 1998 http://xxx.lanl.gov/ps/cmp-lg/9806011

[Arjona-Iglesias, 91] Arjona-Iglesias, M. *Estudios sintácticos sobre el habla popular mexicana*. Universidad Nacional Autónoma de México, 1991.

[Atkins *et al*, 86] Atkins, B., Kegl, J., and Levin, B. *Explicit and Implicit Information in Dictionaries*. In Proceedings of the Second Conference of the UW Center for the New Oxford English Dictionary, pp. 45-65. Waterloo, Canada, 1986.

[Baayen, 92] Baayen, H. *Statistical Models for Word Frequency Distributions: A Linguistic Evaluation*. Computers and the Humanities, 26, pp. 347-363, 1992.

[Baker, 82] *Trainable Grammars for Speech Recognition*. In D. Klatt and J. Wolf (eds.), Speech Communication Papers for the 97th Meeting of the Acoustical Society of America, ASA, pp. 547-550, 1982.

[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, pp. 447-453, 1994.

[Basili, 99] Basili, R., *et al*. *Adaptive parsing and Lexical learning*. In Proceedings Conference Venecia per il Trattamento Automatico delle Lingue (VEXTAL), November 22-24, Venezia, Italy pp. 111- 120, 1999.

[Balkan & Fouvry, 95] L. Balkan and F. Fouvry. *Corpus-based test suite generation*. TSNLP-WP 2.2, University of Essex, 1995.

[Balkan, *et al*, 94] L. Balkan, S. Meijer, D. Arnold, D. Estival, and K. Falkedal. *Test suites for natural language processing.* Translating and the Computer, 16: 51--58, 1994

[Berthouzoz & Merlo, 97] Berthouzoz, C. and Merlo, P. *Statistical ambiguity resolution for principle-based parsing*. In Proceedings of the Recent Advances in Natural Language Processing. Pag. 179-186, 1997

[Biber, 93] Biber, D. Using Register. *Diversified Corpora for general Language Studies*. Computational Linguistics 19 (2) pp. 219—241, 1993.

[Black, 87] Black, D. *The theory of comitees and Elections*. Boston, M. A. Kluwer Academic Press, 1987

[Bleam *et al*, 98] Bleam, T.; Palmer, M. and K. Vijay-Shanker. *Motion verbs and Semantic Features in TAG*. TAG+4 Workshop. University of Pennsylvania, 1998.

[Boguraev & Briscoe, 87] Boguraev, B. and Briscoe, E. *Large lexicons for natural language processing: utilising the grammar coding system of the Longman Dictionary of Contemporary English*. Computational

Linguistics 13.4: 219-240 1987.

[Boguraev *et al*, 87] Boguraev, B., Briscoe, E., Carroll, J., Carter, D. and Grover, C. *The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English*. In Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics, pp. 193-200. Stanford, CA. 1987.

[Bolshakov *et al*, 98] Bolshakov, I., A. Gelbukh, S. Galicia Haro, M. Orozco Guzmán. *Government patterns of 670 Spanish verbs*. Technical report. CIC, IPN ISBN 970-18-1893-8, 1998.

[Bonnema et al, 2000] Bonnema, R., P. Buying, R. Scha. *Parse Tree Probability in Data Oriented Parsing*. In Proceedings International Conference CICLing-2000, February 13-19, Mexico City, pp. 219-232, 2000.

[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.

[Bozsahin, 98] Bozsahin, Cem. *Deriving the Predicate-Argument Structure for a Free Word Order Language*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 167-173, 1998. http://xxx.lanl.gov/ps/cmp-lg/9808008

[Branchadell, 92] Branchadell, A. *A Study of Lexical and Non-lexical datives*. Tesis doctoral inédita. Universitat Autónoma de Barcelona, 1992.

[Brent, 91] Brent, M. *Automatic acquisition of subcategorization frames from untagged text*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 209--214. Berkeley, CA. 1991.

[Brent, 93] Brent, M. *From grammar to lexicon: unsupervised learning of lexical syntax*. Computational Linguistics 19.3: 243--262, 1993

[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

[Bresnan & Kanerva, 88] Bresnan, J. and Kanerva, J. *Locative inversion in Chichewa: A case study of factorization in grammar*. Technical report, Stanford University and Xerox PARC, Stanford, 1988.

[Bresnan & Moshi, 89] Bresnan, J. and Moshi, L. *Object asymmetries in comparative Bantu syntax*, Linguistic Inquiry 21, 1989.

[Brew, 95] Brew, C. *Stochastic HPSG*. In Proceedings of the 7th European Conference of the Association for Computational Linguistics. Pages 83-89, 1995. http://xxx.lanl.gov/ps/cmp-lg/9502022

[Brill, 95] Brill, E. *Unsupervised Learning of disambiguation Rules for Part of Speech Tagging*. In Proceedings of 3rd Workshop on Very Large Corpora. Pages. 1-13. Masachussets, 1995.

[Brill & Resnick, 94] Brill, E., P. Resnik. *A rule-based approach to prepositional phrase attachment disambiguation*. In Proceedings International Conference COLING-94. August 5-9 Kyoto, Japan, pp. 1198-1204, 1994. http://xxx.lanl.gov/ps/cmp-lg/9410026

[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 & Carroll, 97] Briscoe, E. and Carroll, J. *Automatic extraction of subcategorization from corpora*. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing. Washington, DC. 1997. http://xxx.lanl.gov/ps/cmp-lg/9702002

[Briscoe, 96] Briscoe, E. *Robust Parsing*. In The State of the Art of Human Language Technology 1996. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

[Bröker, 2000] N. Bröker. *Improving Testsuites via Instrumentation*. In Proceedings of ANLP--NAACL, Apr 29-May 4, pp. 325-330. 2000. http:// xxx.lanl.gov/ ps/cs.CL/0005016

[Brown *et al*, 90] Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C. and Mercer, R. L. *Class-based n-gram Models of natural Language*. In Proceedings of the IBM Natural language ITL, March, 1990. Paris, France.

[Cano, 87] Cano Aguilar, R. *Estructuras sintácticas transitivas en el español actual*.

Edit. Gredos. Madrid, 1987.

[Carpenter, 95] Carpenter, R. *Categorial Grammars, Lexical Rules and the English Predicative*. Carnegie Mellon University. 1995.

[Carpenter, 97] Carpenter, R. *Type-Logical Semantics*. Cambridge, Mass. MIT Press, 1997.

[Carroll & Rooth, 98] Carroll, G. and Rooth, M. *Valence induction with a head-lexicalized PCFG*. In Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing. Granada, Spain, 1998. http://xxx.lanl.gov /ps/cmp-lg/9805001

[Carroll & Weir, 97] Carroll, J. and Weir, D. *Encoding frequency information in lexicalized grammars*. In Proceedings of the 5th ACL/SIGPARSE International Workshop on Parsing Technologies (IWPT-97), 8--17. MIT, Cambridge, MA. 1997. http://xxx.lanl.gov/ps/cmp-lg/9708012

[Charniak, 93] Charniak, E. *Statistical Language Learning*, MIT, Cambridge, MA. 1993.

[Charniak, 97] Charniak, E. *Statistical parsing with a context-free grammar and word statistics*, Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI MIT Press, Menlo Park, 1997. http://www.cs.brown.edu/ people/ec/home.html#publications

[Chen & Chen, 96] Chen, K. and Chen, H. *A Rule-Based and MT-Oriented Approach to prepositional Phrase Attachment*. In Proceedings of COLING-96, pp. 216-221, 1996.

[Chodorow *et al*, 87] Chodorow, M., Klavans, J., Neff, M., Byrd, R., Calzolari, N. and Rizk, O. *Tools and methods for computational lexicography*. Vol. 13. Pages 3-4. Computational Linguistics, 1987.

[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.

[Chomsky, 95] Chomsky, N. *The Minimalist Program*. Cambridge, Mass. MIT Press, 1995.

[Church & Mercer, 93] Church, K. W. and R. Mercer. *Introduction to the Special Issue on Computational Linguistics Using large Corpora*. 19(1), pp. 1-24, 1993

[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.

[Church *et al*, 91] Church, K. W., Gale, W., Hanks, P., and Hindle, D. *Using statistics in lexical analysis*. In U. Zernik, editor, Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum Associates. Hillsdale, NJ. 1991.

[Civit & Castellón, 98] Civit, M e I. Castellón. *Gramesp: Una gramática de corpus para el español*. Revista de AESLA, La Rioja, España, 1998.

[Collins, 99] Collins, M. *Head-driven Statistical Models for Natural language parsing*. Ph.D. Thesis University of Pennsylvania. 1999. http:// xxx.lanl.gov/ find/cmp-lg/

[Collins & Brooks, 95] Collins, M. and J. Brooks. *Prepositional phrase attachment through a backed-off model*. In Proceedings of the 3rd Workshop on Very Large Corpora. Pag. 27--38. Cambridge, MA. USA, 1995 http:// xxx.lanl.gov/ps/cmp-lg/9506021

[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.

[Debreu, 59] Debreu, G. *The Theory of Value: An axiomatic analysis of economic equilibrium*, 1959 Theory of Value : An Axiomatic Analysis of Economic Equilibrium. Yale University Press, 1986.

[DECIDE, 96] The DECIDE project. *Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora*. 1996. http:// engdep1.philo.ulg.ac.be/ decide/

[Demonte, 94] Demonte, V. *La ditransividad en español: léxico y sintaxis*, en Gramática del Español. Edición a cargo de Violeta Demonte El Colegio de México, 1994.

[DEUM, 96] DEUM *Diccionario del Español Usual en México*. Edit. Colegio de México. México, 1996.

[DG Website, 99] DG Website *Dependency-Based Approaches to Natural Language Syntax*. 1999. http://ufal.mff.cuni.cz/dg/ dgmain.html

[Dowty, 82] Dowty, D. R. *Grammatical relations and Montague Grammar*. In P. Jacobson and G. K. Pullum, eds., the Nature of Syntactic Representation, pps. 79--130, Reidel, Dordrect, 1982.

[Dowty, 89] Dowty, D. R. *On the semantic content of the notion "thematic role"*. In G. Chierchia, B. Partee and R. Turner (eds). Property theory, type theory and natural language semantics. D. Reidel, Dordrecht, 1989.

[Dowty, 91] Dowty, D. R. *Thematic proto-roles and argument selection*. Language 67.3, 547-619 1991.

[EAGLES, 96] EAGLES. *Recommendations on Sub-categorization*, 1996. http://www.ilc.pi.cnr.it/ EAGLES96/ syn-lex/ synlex. html

[Edmundson, 63] Edmundson, H. P. *A Statistician's View of Linguistic Models and Language Data Processing*, in Garvin, P. L. (ed.), Natural Language and the Computer, new York: McGraw Hill, 1963.

[Eisner, 96] Eisner, J. M. Three New Probabilistic Models for Dependency Parsing: An Exploration. In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96). Pages 340--345, 1996. http://xxx.lanl.gov/ps/cmp-lg/9706003

[Elworthy, 94] Elworthy, D. *Does Baum-Welch re-estimation help taggers?*. In Proceedings of the 4th Conference of Applied NLP. Stuttgart, Germany, 1994. http://xxx.lanl.gov/ps/cmp-lg/9410012

[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

[Fabre, 96] Fabre, C. *Recovering a predicate-Argument Structure for the Automatic Interpretation of English and French Nominal Compounds*. In Proceedings of the International Workshop on Predicative Forms in

Natural Language and in Lexical Knowledge Bases. 27--34. Toulouse, France, 1996.

[Fillmore, 68] Fillmore, C. J. *The case for case*. In: E. W. Bach and R. T. Harms (eds.). Universals in Linguistic Theory. New York: Holt, Rienhart & Winston, 1968.

[Fillmore, 71] Fillmore, C. J. *Types of lexical information*. In Semantics: an interdisciplinary reader, pp. 370-392. Steinberg & Jakobovits. Cambridge University Press, 1971.

[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.

[Fishburn & Gehrlein, 76] Fishburn, P. C. and Gehrlein, W. V. *Borda's rule, Positional Voting, and Condorcet's Simple Majority Principle*. Public Choice, Vol. 28. Pp-79-88, 1976.

[Flickinger, *et al*, 85] Flickinger, D., C. Pollard, and T. Wasow. *"Structure sharing in lexical representation,"* Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics. Chicago, IL: Association for Computational Linguistics, 262-7, 1985.

[Flickinger *et al*, 87] D. Flickinger, J. Nerbonne, I. Sag, and T. Wasow. *Toward Evaluation of NLP Systems*. Hewlett-Packard Laboratories, Palo Alto, CA., 1987.

[Ford *et al*, 82] Ford, M., Bresnan, J. and Kaplan, R. *A competence based theory of syntactic closure*. In Bresnan, J. W., editor, The Mental Representation of Grammatical Relations, MIT Press, Cambridge, MA. 1982.

[Franks, 69] Franks, L. E. *Signal Theory.* Prentice Hall, Englewood Cliffs, N.J. 1969

[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.

[Fujisaki *et al*, 89] Fujisaki, T., Jelinek, F., Cocke, J., and Black, E. *Parsing, word associations and typical predicate-argument relations.* In Proceedings of

the International Workshop on Parsing Technologies Carnegie-Mellon University, 1989

[Gale *et al*, 92] Gale, W. A., Kenneth W. Church, and David Yarowsky. *Work on Statistical Methods for Word Sense Disambiguation*. In Probabilistic Approaches to Natural Language: Papers from the 1992 Fall Symposium, pp. 54-60, Cambridge, Massachusetts. Menlo Park, Calif. American Association for Artificial Intelligence, AAAI Press, 1992.

[Galicia *et al*, 97] Galicia, Haro Sofía N., Alexander F. Gelbukh, Igor A. Bolshakov. *Patrones de manejo sintáctico para verbos comunes del español*. In Proceedings CIC-97, Nuevas Aplicaciones e Innovaciones Tecnológicas en Computación, Simposium Internacional de Computación, November 12-14, 1997, CIC, IPN, Mexico City, Mexico.

[Galicia *et al*, 98] Galicia Haro, S., I. Bolshakov, A. Gelbukh. *Diccionario de patrones de manejo sintáctico para análisis de textos en español*. Revista SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural, No. 23, España, pp.171-176. Septiembre de 1998.

[García-Hidalgo, 79] García-Hidalgo, M. I. *La formalización del analizador gramatical del DEUM*. En Lara, L. F.; Ham Chande, R. y García Hidalgo, M. I. Investigaciones lingüísticas en Lexicografía. El Colegio de México, 1979.

[Gawron *et al*, 82] Gawron, J. M.; King, J.; Lamping, J.; Loebner, J.; Paulson, E.; Pullum, J.; Sag, I.; and Wasow, T. *Processing English with a generalized phrase structure grammar*. In Proceedings ACL, pp. 74-81, 1982

[Gazdar *et al*, 85] Gazdar, G., E. Klein, G. K. Pullum, and I. A. Sag. *Generalized Phrase Structure Grammar*. Oxford, Blackwell, 1985.

[Gee & Grosjean, 83] Gee, James Paul and François Grosjean. *Performance structures: A psycholinguistic and linguistic appraisal*. Cognitive Psychology (15): 411--458, 1983.

[Gelbukh *et al*, 98] Gelbukh, A., S. Galicia-Haro, I. Bolshakov. *Three dictionary-based techniques of disambiguation*. In Proceedings TAINA-98, International Workshop on Artificial Intelligent, CIC-IPN, Mexico D. F., pp. 78 - 89, 1998.

[Gelbukh, 98] Gelbukh, A. F. *Lexical, syntactic, and referencial disambiguation using a semantic network dictionary*. Technical report. CIC, IPN, 1998.

[Gibbon, 99] Gibbon, D. *Computational lexicography*. ELSNET Group 1999. http://coral.lili.uni-bielefeld.de/~gibbon/ELSNET97/index.html

[Gili, 61] Gili Gaya, S. *Curso Superior de Sintaxis Española*. Bibliograf. España, 1961.

[Goodman, 98] Goodman, Joshua T. *Parsing inside-out*. Ph. D. thesis, Harvard University, Cambridge, MA. http:// xxx.lanl.gov/abs/cmp-lg/9805007.

[Greffenstette, 93] Greffenstette, G. *Evaluation techniques for automatic semantic extraction: Comparing syntactic and window based approaches*. In ACL Workshop on Acquisition of Lexical Knowledge From Text, Ohio State University, June, 1993.

[Grimes, 72] Grimes, J. E. *The thread of discourse*. Technical report NSF. Cornell University, 1972.

[Grimes, 75] Grimes, J. E. *The thread of discourse.* Moulton Press, The Hague, 1975.

[Grinberg *et al*, 95] Grinberg, D., Lafferty, J. and Sleator, D. *A Robust parsing Algorithm for Link grammars*. In Proceedings of the Fourth International Workshop on Parsing Technologies. Pag. 111—125, 1995. http:// xxx.lanl.gov/ps/cmp-lg/9508003

[Grishman *et al*, 94] Grishman, R., Macleod, C. and Meyers, A. *COMLEX syntax: building a computational lexicon*. In Proceedings Conference COLING-94, Kyoto, Japan, pp. 268-272 1994. http://xxx.lanl.gov/ps/cmp-lg/9411017

[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. http://www.gs.uni-heidelberg.de/~hellwig/

[Hellwig, 83] Hellwig, P. *Extended Dependency Unification Grammar*. In: Eva Hajicova (ed.): Functional Description of Language. Faculty of Mathematics and Physics, Charles University, Prague, pp. 67-84, 1983. http: //www.gs.uni-heidelberg.de/~hellwig/biblio.html

[Hellwig, 86] Hellwig, P. *Dependency Unification Grammar (DUG)*. In: Proceedings of the 11th International Conference on Computational Linguistics (COLING 86), 195-198. Bonn: Universität Bonn, 1986. http://www.gs.uni-heidelberg.de/~hellwig/

[Hellwig, 95] Hellwig, P. *Automatic Syntax Checking*. In: M. Kugler, K. Ahmad, G. Thurmair (eds.): Translator's Workbench. Berlin, Heidelberg, New York: Springer, 1995. http://www.gs.uni-heidelberg.de/~hellwig/

[Hellwig, 98] Hellwig, P. *Parsing - A Course in Cooking*. Tutorial given at the COLING/ACL-98 conference in Montreal. Tutorial Notes, 1998. http://www.gs.uni-heidelberg.de/~hellwig/biblio.html

[Hindle & Rooth, 93] Hindle, D. and M. Rooth. *Structural ambiguity and lexical relations*. Computational Linguistics, 19(1): 103--120, 1993

[Holmes, 88] Holmes, J. *Speech Synthesis and Recognition*. Van Nostrand Reinhold. Workingham, UK. 1988.

[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 Contemporarary Research. Berlin: Walter de Gruyter. 1998. http://www.phon.ucl.ac.uk/home/dick/ wg.htm

[Ilson & Mel'cuk, 89] Ilson, R. and I. A. Mel' cuk. *English BAKE Revisited (BAKE-ing an ECD).* International Journal of Lexicography, 2(4), 326-45. 1989.

[Jackendoff, 90] Jackendoff, R. S. *Semantics Structures*. MIT Press, Cambridge, MA. 1990.

[Jacobs *et al*, 91] Jacobs, P. S., Krupka, G. R., and Rau, L. F. *Lexico-semantic pattern matching as a companion to parsing in text understanding.* In Proceedings of the Fourth DARPA Speech and natural language Workshop, pp. 337-342. Pacific Grove, CA, USA. 1991.

[Jelinek *et al*, 92] Jelinek, F., R. L. Mercer, and S. Roukos. *Principles of Lexical Language Modeling for Speech Recognition*. In S. Furui and M. M. Sondhi, editors, Advances in Speech Signal Processing. New York, Marcel Dekker, 1992.

[Jiang & Conrath, 97] Jiang, J. and Conrath, D. *Semantic Similarity on Corpus Statistics and Lexical Taxonomy*. In Proceedings of the 10th International

Conference Research on Computational Linguistics (ROCKLING'97). Taiwan, 1997. http://xxx.lanl.gov/ps/cmp-lg/9709008

[Jones, 94] Jones, B. E. M. *Exploring the Role of Punctuation in Parsing Natural Text*. In Proceedings International Conference COLING-94. August 5-9 Kyoto, Japan, pp. 421-425, 1994. http://xxx.lanl.gov/ps/cmp-lg/9505024

[Joshi, 85] Joshi, A. K. *How much Context-Sensitivity is Necessary for Characterizing Structural Descriptions - Tree Adjoining Grammars*. In Dowty, D.; Karttunen, L.; and Zwicky, A. (eds.), Natural language Processing - Theoretical, Computational and Psychological Perspectives. Cambridge University Press, New York, 1985.

[Kahn, 66] Kahn, D. *The Codebreakers*, London. Weidenfeld and Nicolson, 1966.

[Kaplan, 94] Kaplan, R. M. *The Formal Architecture of Lexical Functional Grammar*. In: Dalrymple, M., *et al*. (eds.). Formal Issues in Lexical Functional Grammar. Stanford University Press, 1994.

[Kaplan & Bresnan, 82] Kaplan, R. M. and Bresnan, J. *Lexical functional grammar: A formal system for grammatical representation*. In J. Bresnan, editor, The Mental Representation of Grammatical Relations. Pages 173-281. MIT Press, Cambridge, MA. 1982.

[Karlsson *et al*, 95] Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. *Constraint grammar: A Language-Independent System for Parsing Unrestricted text*, edited by authors. Mouton de Gruyter, 1995.

[Kasami, 65] Kasami, J. *An efficient recognition and syntax analysis algorithm for context-free languages*. Technical Report. University of Hawaii. 1965

[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, pp. 155-188.

[Kay, 80] Kay, M. *Algorithm Schemata and data Structures in Syntactic processing*. Report CSL-80-12, Xerox PARC, Palo Alto, CA. 1980. Reprinted in: Grosz, B. J. *et al* (eds.), Readings in Natural language Processing. Morgan Kaufmann, Los altos, CA. 1982.

[Kilgarriff, 92] Kilgarriff, A. *Polysemy*. PhD thesis,University of Sussex, CSRP 261, School of Cognitive and Computing Sciences, 1992.

[Kilgarriff, 93] Kilgarriff, A. *Inheriting Verb Alternations*. In Proceedings 6th European Conference of ACL. pp. 213-221. Utrecht, Netherlands, 1993.

[Kittredge, 2000] Kittredge, R. *Interlingual Modelling: An Applications Perspective*. In Proceedings International Conference CICLing-2000, February 13-19, Mexico City, pp. 19-29, 2000.

[Krause & Clark, 93] Krause, P. and Clark, D. *Representing Uncertain Knowledge*. Dordrecht, Kluwer Academic Publishers, 1993.

[Kupiec, 91] Kupiec, J. *A trellis-based algorithm for estimating the parameters of a hidden stochastic context-free grammar*. In Proceedings of the DARPA Speech and Natural Language Workshop, Hidden Valley, Penn, 1991

[Lamiroy, 94] Lamiroy, B. *Causatividad, ergatividad y las relaciones entre el léxico y la gramática*. En Gramática del Español, edición a cargo de Violeta Demonte. El Colegio de México, 1994.

[Lansdowne, 96] Lansdowne, Z. F. *Ordinal ranking Methods for Multicriterion Decision Making*. Naval Research Logistics, Vol. 43, pp. 613-627, 1996. Reprinted in MITRE Journal, pp. 23-36, 1997.

[Lara & Ham, 79] Lara, L. F. y Ham Chande, R. *Base estadística del Diccionario del español de México*. En Lara, L. F.; Ham Chande, R.; García Hidalgo, M. I. (eds.) Investigaciones lingüísticas en Lexicografía. El Colegio de México 1979.

[Lari & Young, 90] Lari, K. and S. Young. *The estimation of stochastic context-free grammars using the Inside-Outside Algorithm*, Computer Speech and Language Processing, vol.4, pp. 35-56, 1990.

[Leech & Garside, 91] Leech, G. and R. Garside. *Running a grammar factory: the production of syntactically analysed corpora or treebanks*. In S. Johansson and A. Stenstrom, English Computer Corpora: Select Berlin, 1991

[Levin, 93]Levin, B. *English Verb Classes and Alternations*. University of Chicago Press, 1993.

[Levin & Rappoport, 91] Levin, B. and Rappoport Hovav, M. *Wiping the slate clean: A lexical semantic exploration*. Cognition, 41: 123- 151, 1991.

[Litkowski, 92] Litkowski, K. C. *A primer on computational lexicology*. 1992. http://www.clres.com/

[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, pp. 787-793, 1998.

[Ludwig, 96] Ludwig, B. *POS Tagging Using Morphological Information*. 1996. http://xxx.lanl.gov/ ps/cmp-lg/9606005

[Luna-Traill, 91] Luna-Traill, E. *Sintaxis de los verboides en el habla culta de la Ciudad de México*. Universidad Nacional Autónoma de México, 1991.

[Lyons, 77] Lyons, J. *Semantics*. Cambridge, 1977.

[Magerman, 95] Magerman, D. M. *Statistical decision-Tree Models for Parsing*. In Proceedings 33rd Annual Meeting of ACL. June 26-30 Cambridge, Massachusetts, USA, pp. 276-283, 1995. http://xxx.lanl.gov/ps/cmp-lg/9504030

[Manning & Carpenter, 97] Manning, C. and B. Carpenter. *Probabilistic parsing using left corner language models*. In Proceedings of the 5th Intl. Workshop on Parsing Technologies, 1997. http://xxx.lanl.gov/ps/cmp-lg/9711003

[Manning, 93] Manning, C. *Automatic acquisition of a large subcategorization dictionary from corpora*. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, 235--242. Columbus, Ohio, 1993.

[Marcus *et al*, 93] Marcus, M., Santorini, B. and Marcinkiewicz, M. *Building a large annotated corpus of English The Penn Treebank*. Computational Linguistics 19, 2, 1993.

[Marcus *et al*, 94] Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, M., Ferguson, M., Katz, K. and Schasberger, B. *The Penn Treebank: Annotating predicate argument structure*. In Proceedings Human Language Technology Workshop, Morgan Kaufmann, San Francisco, 1994

[Markov, 16] Markov, A. A. *An Application of Statistical Method*. Izvestiya Imperialisticheskoj akademii nauk, 6(4), pp. 239-42, 1916.

[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 & Zholkovsky, 84] Mel'cuk, I. A. and A. K. Zolkovsky. *Explanatory combinatory dictionary of modern Russian*. Wiener Slawistischer Almanach, Vienna, 1984.

[Mel'cuk *et al*, 84] Mel'cuk, I. A., N. Arbatchewsky-Jumarie, L. Elnitsky, *et al*. *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques I*. Presses de l'Université de Montréal, Montreal, 1984

[Mel'cuk *et al*, 88] Mel'cuk, I. A., N. Arbatchewsky-Jumarie, L. Dagenais, *et al*. *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques II*. Presses de l'Université de Montréal, Montreal, 1988

[Mel'cuk & Pertsov, 87] Mel'cuk, I. A., and Nikolaj V. Pertsov. *Surface Syntax of English: a Formal Model within the Meaning-Text Framework*. Amsterdam, Benjamins, 1987

[Mel'cuk, 88] Mel'cuk, I. *Dependency Syntax: Theory and Practice*, New York: State University of New York Press, 1988.

[Mel'cuk, 88a] Mel'cuk, I. *Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria*. International Journal of Lexicography Vol. 1, No. 3, pp.165-188, 1988.

[Merlo et al, 97] Merlo, P., Crocker, M. and Berthouzoz, C. *Attaching Multiple Prepositional Phrases: Generalized Backed-off Estimation.* In Proceedings of the EMNLP-2, 1997. http://xxx.lanl.gov/find/cmp-lg/9710005

[Meyer *et al*, 90] Meyer, I., B. Onyshkevych, and L. Carlson. *Lexicographic Principles and Design for Knowledge-Based Machine Translation*, Technical Report CMU-CMT-90-118. Pittsburgh, PA: Carnegie Mellon University, Center for Machine Translation, 1990.

[Miller, 90] Miller G. *Wordnet: an on-line lexical database*. International Journal of Lexicography, 3(4), 1990.

[Mohri & Pereira, 98] Mohri. Mehryar, F. C. N. Pereira. *Dynamic Compilation of Weighted Context-free Grammar*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 891-897, 1998.

[Monedero *et al*, 95] Monedero, J., González, J. C., Goñi, J. M., Iglesias, C. A. y Nieto, A. *Obtención automática de marcos de subcategorización verbal a*

partir de texto etiquetado: el sistema SOAMAS. En Actas del XI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 95: Bilbao), págs. 241-254, 1995.

[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.

[Moortgat, 94] Moortgat M., Oehrle R. *Adjacency, dependency and order*. P. Dekker and M. Stokhof, eds. In Proceedings of the 9th Amsterdam Colloquium. Dordrecht: Foris, 1994.

[Moreno, 85] Moreno de A., J. G. *Valores de las formas verbales del español de México*. Universidad Nacional Autónoma de México, 1985.

[Mueller, 96] Mueller, Dennis C. (ed.) *Perspectives on Public Choice*. A Handbook. Cambridge University Press, 1996.

[Nañez, 95] Nañez Fernández, E. *Diccionario de construcciones sintácticas del español*. Preposiciones. Ed. de la Universidad Autónoma de Madrid, España 1995.

[Netter *et al*, 98] K. Netter, S. Armstrong, T. Kiss, J. Klein, and S. Lehman. *Diet - diagnostic and evaluation tools for NLP applications*. In Proceedings 1st International Conference on Language Resources and Evaluation, pages 573--579. Granada/Spain, 28-30 May, 1998.

[Osborne, 96] Osborne, M. *Can Punctuation Help Learning?*. In Connectionist, statistical and Symbolic Approaches to Learning for Natural Language Processing, edited by S. Wermter, E. Riloff and G. Scheler. Springer Verlag, 1996.

[Padró, 98] Padró, L. *A Hybrid Environment for Syntax-Semantic Tagging*. Ph. D. Thesis. Departament de Llenguatges I Sistemes Informàtics de la Universitat Politècnica de Catalunya. 1998. http://xxx.lanl.gov/ps/cmp-lg/9802002

[Pedersen, 2000] Pedersen, T. *An ensemble Approach to Corpus-based Word Sense Disambiguation*. In Proceedings International Conference CICLing-2000, February 13-19, Mexico City, pp. 205-218. 2000.

[Penadés, 94] Penadés Martínez, I. *Esquemas Sintáctico-Semánticos de los Verbos Atributivos del Español*. Servicio de Publicaciones. Universidad de Alcalá. España, 1994.

[Pereira, 96] Pereira, F. *Sentence Modelling and Parsing*. In The State of the Art of Human Language Technology. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html 1996

[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.

[Picinbono, 80] Picinbono, B. C. *A geometrical interpretation of signal detection and estimation*. IEEE Transactions of Information Theory, 26(4), pp. 493-497. 1980

[Pirelli *et al*, 94] Pirelli, V., N. Ruimy, and S. Montemagni. *Lexical regularities and lexicon compilation*. Acquilex-II Working Paper 36, 1994. http://www.cl.cam.ac.uk/Research/NL/Acquilex/

[Polguère, 98] Polguère, A. *Observatory of Meaning-Text Linguistics (OMTL)*. Université de Montréal. Faculté des Arts et des Sciences. 1998. http://www.fas.umontreal.ca/ling/olst/indexE.html

[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.

[Pratt, 42] Pratt, F. *Secret and Urgent*. Garden City, NY; Blue Ribbon Books, 1942.

[Procter *et al*, 78] Procter, P. *et al*. *Longman Dictionary of Contemporary English (LDOCE)*. Longman Group, Harlow, Essex, UK. 1978.

[Procter, 87] Procter, P. *Longman Dictionary of Contemporary English*, Longman, London, 1987.

[Rabiner, 89] Rabiner, L. *A tutorial in hidden Markov models and selected applications in speech recognition*. In proceedings IEEE, 77(2), 257-286, 1989.

[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.

[Rappaport, 83] Rappaport, M. *On the nature of derived nominals*. In L. Levin, M. Rappaport and A. Zaenen (eds), Papers in Lexical-Functional Grammar, University Linguistics Club, Bloomington, Indiana, 1983.

[Ratnaparkhi *et al*, 94] Ratnaparkhi, A., J. Reynar, and S. Roukos. *A maximum entropy model for prepositional phrase attachment*. In Proceedings of the Human Language Technology Workshop. Advanced Research Projects Agency, March, 1994.

[Ratnaparkhi, 98] Ratnaparkhi, A. *Statistical Models for Unsupervised Prepositional Phrase Attachment*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Quebec, Canada, 1998 http://xxx.lanl.gov/ps/cmp-lg/9807011

[Resnik & Hearst, 93] Resnick, P. and Hearst, M. *Syntactic ambiguity and conceptual relations*. In: K. Church (ed.) Proceedings of the ACL Workshop on Very Large Corpora, pp. 58-64, 1993

[Rieger & Small, 82] Rieger, C. and Small, S. *Parsing and comprehending with word experts (a theory and its realisation)*. In Strategies for Natural Language Processing, Lehnert and Ringle, Erlbaum, Hillsdale, N. J., pp. 89 - 147, 1982.

[Rigau *et al*, 97] Rigau, G., Atserias, J. and Agirre, E. *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 48-55, 1997 http://xxx.lanl.gov/ps/cmp-lg/9704007

[Rodríguez *et al*, 98] Rodríguez, H., Climent, S., Vossen, P., Blocksma, L., Peters, W., Alonge, A., Bertagna, F. and Rovertini, A. *The top-down strategy for buiding EUWN: Vocabulary coverage, base concepts and Top-Ontology*. In N. Ide and D. Greenstein (eds.) Computers and the humanities, vol. 32, n. 2-3, 1998.

[Rojas, 88] Rojas, C. *Verbos locativos en español*. Aproximación sintáctico-semántica. Universidad Autónoma de México, 1988.

[Roland & Jurafsky, 98] Roland. D. and D. Jurafsky. *How Verb Subcategorization*

Frequencies are Effected by Corpus Choice. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1122-1128, 1998.

[Rosenfeld, 94] Rosenfeld, R. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach.* Ph.D. thesis, Computer Science Department, Carnegie Mellon University, 1994. http://www.cs.cmu.edu/ afs/cs.cmu.edu/user/roni/ WWW/ HomePage.html

[Saari, 94] Saari, D. G. *Geometry of Voting*. NewYork, Springer-Verlag, 1994

[Sag & Wasow, 99] Sag, I. A. and Wasow, T. *Syntactic Theory: A Formal Introduction*. Center for the study of language and information, 1999.

[Salomaa, 71] Salomaa, A. *The generative power of transformational grammars of Ginsburg and Partee*. Information and Control, 18, pp. 227-232, 1971.

[Samuelson & Voutilainen, 97] Samuelson, C, and A. Voutilainen. *Comparing a linguistic and a Stochastic tagger*. In Proceedings of joint ACL/EACL, Madrid, Spain, 1997.

[Sanfilippo & Poznanski, 92] Sanfilippo, A. and Poznanski, V. *The acquisition of lexical knowledge from combined machine readable dictionary sources*, Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, 1992.

[Sanfilippo, 93] Sanfilippo, A. *LKB encoding of lexical knowledge*. In T. Briscoe, A. Copestake and V. De Paiva (eds.). Default inheritance in unification-based approaches to the lexicon. CUP. Cambridge, 1993

[Sanfilippo, 97] Sanfilippo, A. *Using Similarity to Acquire Coocurrence Restrictions from Corpora*. pp. 82-87, 1997

[Santorini, 90] Santorini, B. *Penn Treebank Tagging and Parsing Manual*, University of Pennsylvania, CIS Dept. 1990.

[Scharf, 91] Scharf, L. L. *Statistical Signal Processing*. Addison Wesley, Reading MA. 1991.

[Schütze & Gibson, 99] Schütze, C. T. and Gibson, E. *Argumenthood and English prepositional Phrase Attachment*. Journal of Memory and Language, vol. 40, n. 3, pp. 409-431, 1999

[Schabes, 92] Schabes, Y. *Stochastic lexicalized tree-adjoining grammars*. In Proceedings of the 14th International Conference on Computational

Linguistics (COLING-92), 426--432. Nantes, France, 1992

[Schank *et al*, 72] Schank, R., Neil Goldman, Charles Rieger and Christopher Riesbeck. *MARGIE: Memory, analysis, response generation and inferences in english*. IJCAI Actes 3º, 1972.

[Schank, 80] Schank, R., M. Lebowitz and Lawrence Birnbaum. *An integrated understander*. American Journal of Computational Linguistics 6 (1), 1980.

[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, pp. 104-110, 1992.

[Sells, 85] Sells, P. *Lectures on Contemporary Syntactic Theories*. CSLI Lecture Notes, Stanford, CA. Number 3, 1985.

[Shannon, 49] Shannon, C. E. *The Mathematical Theory of Communication*, in Shannon, C. and Weaver, W. (eds.), The Mathematical Theory of Communication. Urbana, IL. The University of Illinois Press, 1949.

[Sharman, 89] Sharman, R. A. *An introduction to the Theory of Language Models*, IBM UKSC Report 204 1989.

[Shieber, 84] Shieber, S. M. *An introduction to Unification-Based Approaches to Grammar*. Stanford, Calif., Center for the Study of Language and Intonation, 1986.

[Shtrikman, 94] Shtrikman, S. *Some Comments on Zipf's Law for the Chinese Language*. Journal of Information Science, 20 (2), pp. 142-3, 1994.

[Sikkel & Akker, 93] Sikkel, K. and R. op den Akker. *Predictive Head-Corner Chart Parsing*. In Procedings of the 3rd International Workshop on Parsing Technologies (IWPT'3), pages 267--276, 1993.

[Sikkel, 97] Sikkel, Klass. *Parsing Schemata*. Springer, 1997.

[Sinclair *et al*, 87] Sinclair, J. M., Hanks, P., Fox, G., Moon, R., and Stock, P., editors. *Collins Cobuild English Language Dictionary Collins*, London, 1987

[Sleator & Temperley, 93] Sleator, D. and D. Temperley. *Parsing English with a link grammar*. In 3rd International Workshop on Parsing Technologies (IWPT'3), pages 277--292, 1993

[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.

[Steele, 90] Steele, J. *Meaning - Text Theory. Linguistics, Lexicography, and Implications*. James Steele, editor. University of Ottawa press, 1990.

[Suppes, 70] Suppes, P. *Probabilistic grammars for natural languages*. Vol. 22. Pages 95-116. Syntheses, 1970

[Tapanainen *et al*, 97] Tapanainen, P., Järvinen, T., Heikkilä, J., Voutilainen. A. *Functional Dependency Grammar*. 1997. http://www.ling.helsinki.fi/~tapanain/dg/

[Tesnière, 59] Tesniere, L. *Elements de syntaxe structural*. Paris: Klincksiek. (German: Tesniere, L. (1980): Grundzüge der strukturalen Syntax. Stuttgart: Klett-Cotta.) 1959.

[Tomita, 86] Tomita, M. *Efficient Parsing for Natural Language*. Kluwer Publ., 1986

[Tzoukerman *et al*, 94] Tzoukerman, E., Radev, D. R. and Gales, W. A. *Combining Linguistic Knowledge and Statistical learning in French Part-of-Speech Tagging*. In Proceedings of the EACL-SIGDAT Workshop From texts to tags. Issues in Multilingual Language Analysis. Pages. 51- 57. Dublin, Ireland, 1994

[Ushioda *et al*, 93] Ushioda, A., Evans, D., Gibson, T. and Waibel, A. *Frequency Estimation of verb Subcategorization Frames Based on Syntactic and Multidimensional Statistical Analysis*. In Proceedings of the Third International Workshop on Parsing Technologies. Pages. 309--318, 1993

[Utsuro, 98] Utsuro, Takehito *et al*. *General-to-Specific Model Selection for Subcategorization Preference*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1314-1320, 1998.

[Uszkoreit & Zaenen, 96] Uszkoreit, H. and Zaenen, A. *Grammar Formalisms*. In The State of the Art of Human Language Technology. 1996 http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

[Uszkoreit, 96] Uszkoreit, H. *Mathematical Methods: Overview*. In The State of the Art of Human Language Technology, 1996.

[Van Newenhizen, 92] Van Newenhizen, J. *The Borda method is Most Likely to Respect the Condorcet Principle*. Economic Theory, vol. 2, pp. 69-83, 1992

[Vanocchi *et al*, 94] Vanocchi, M., Rosini, R., Carenini, M., Prodanof, I. and Calzolari, N. *Italian verbs: Developing a neutral formalism for verbal representation*, Technical Report ILC-NLP-1994-1, ILC-CNR, Pisa, 1994.

[Volk, 92] Volk, M. *The Role of testing in Grammar Engineering*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, pp. 257- 258, 1992.

[Voutilainen, 94] Voutilainen, A. *Three studies of grammar-based surface parsing of unrestricted English text*. Ph. D. Thesis. Department of General Linguistics, University of Helsinki, Finland. 1994. http:// xxx.lanl.gov/ps/ cmp-lg/9406039

[Voutilainen, 95] Voutilainen, A. *Morphological disambiguation*. In Constraint grammar Edited by F. Karlsson, A. Voutilainen, J. Heikkilä and A. Anttila. Pp. 165 -284, 1995.

[Walker *et al*, 95] Walker, D. E., Zampolli, A., and Calzolari, N. *Automating the Lexicon*: Research and Practice in a Multilingual Environment, Oxford University Press, 1995.

[Weaver, 49] Weaver, W. *Recent Contributions to the Mathematical Theory of Communication*, in Shannon, C. and Weaver, W. (eds.), The Mathematical Theory of Communication. Urbana, IL. The University of Illinois Press 1949.

[Whittemore *et al*, 90] Whittemore, G., Ferrara, K. and Brunner, H. *Empirical Study of Predictive Powers of Simple Attachment Schemes for Post-modifier Prepositional Phrases*. In Proceedings of the 28th Annual meeting of the Association for Computational Linguistics. Pages 23-30, 1990

[Wilkins, 97] Wilkins, W. *El lexicón posminimista: el caso SE*. En Estudios de lingüística formal. Pags. 67 - 86. El Colegio de Mexico. México, 1997.

[Williams, 80] Williams, E. *Predication*, Linguistic Inquiry 11: 81-114, 1980.

[Williams, 81] Williams, E. *Argument Structure and Morphology*. Linguistic Review, No. 1, pp. 81-114, 1981.

[Wood, 93] Wood, M. *Categorial Grammars*. Linguistic Theory Guides. London and New York: Routledge, 1993.

[XTAG, 95] XTAG *A Lexicalized Tree Adjoining Grammar for English*. The XTAG Research Group. Technical Report (IRCS-95-03), 1995. http:// www.cis.upenn.edu/ ~cliff-group/94/xtag.html

[Yarowsky, 92] Yarowsky, D. *Word sense disambiguation using statistical models of Roget's categories trained on a large corpus*. In Proceedings of the COLING-92, Nantes, Fr.,pp. 454-460, 1992

[Yarowsky, 95] Yarowsky, D. *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189-196, 1995

[Yeh & Vilain, 98] Yeh, Alexander S., M. B. Vilain. *Some Properties of Preposition and Subordinate Conjunction Attachments*. In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1436-1442, 1998. http://xxx.lanl.gov/ps/cmp-lg/9808007

[Younger, 67] Younger, D. H. *Recognition and parsing of context-free languages in time* n^3. Information and Control 10, pp.189- 208, 1967.

[Yuret, 98] Yuret, D. *Discovery of Linguistic Relations Using Lexical Attraction*. Ph. D. thesis. Massachusetts Institute of Technology. 1998. http:// xxx.lanl.gov/ find/cmp-lg/9805009

[Zipf, 35] Zipf, G. K. *The Psicho - Biology of Language*. Boston, MA. Houghton Mifflin, 1935.

[Zipf, 49] Zipf, G. K. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Publishers Co., Reading, MA. 1949.

[Zubizarreta, 94] Zubizarreta, M. L. *El orden de palabras en español y el caso nominativo*, en Gramática del Español. Edición a cargo de Violeta Demonte El Colegio de México, 1994.

APÉNDICE
CONJUNTO DE PRUEBA

¡ Llamaré a la policía!

Decidió que haría pintar la casa.

Pero Irene la detuvo con un gesto.

No le hace mal a nadie - sonrió.

Beatriz abandonó su puesto de observación mordiéndose los labios.

Este negocio no ha resultado ninguna maravilla.

Voy a entrevistar una especie de santa.

Dicen que hace milagros.

Beatriz suspiró sin dar muestras de apreciar el humor de su hija.

Tenía el hábito de hablar con Dios.

¿ No podía hacerlo en silencio y sin mover los labios ?

Así sucedía en todas las familias.

No quería dar la impresión de haberla descuidado , porque la gente murmuraría a sus espaldas.

Era un período de reposo , descansaban los campos , los días parecían más cortos , amanecía más_tarde.

Siempre lo dijo , pero nadie le prestó atención.

Tenía un carácter galante.

Al volver los hombres el hecho estaba consumado y debieron aceptarlo.

Luego la envió de regreso a su cama.

¿ Qué pensaría su marido al verla ?

Marchaba a su lado con paso firme en las manifestaciones callejeras.

En íntima colaboración criaron a sus hijos.

Esa criatura rubia de ojos claros tal_vez significaba algo en su destino.

Por allí dicen que se comprarán un tractor.

Aunque vivían a escasa distancia tenían pocas ocasiones de encontrarse , pues sus vidas eran muy aisladas.

Cumplía múltiples ocupaciones bajo la tienda.

Ella también lo prefería así.

Su mujer nunca pudo recibirlo con naturalidad.

A_diferencia_de otros campesinos , se casaron enamorados y por amor engendraron hijos.

Nada se botaba ni perdía.

Nada podemos hacer.

Su madre recordaba con exactitud el comienzo de la desgracia.

Entretanto los batracios formaron filas compactas y emprendieron marcha ordenadamente.

La crisis duró pocos minutos y dejó a Evangelina extenuada , a la madre y al hermano aterrorizados.

Nos vamos a arruinar.

Pero todo había sido en_vano.

En su presencia se sentía repudiado.

El joven parecía tener las ideas claras y éstas no coincidían con las suyas.

En ese sentido era muy cuidadosa.

Sus abundantes batallas fortalecieron el odio.

Dejaron la perra en la casa , subieron en la motocicleta y partieron.

Apretaban los dientes y aguantaban callados.

Sacó por fin la voz y se presentó.

Poco después apareció Irene_Beltrán y pudo verla de cuerpo entero.

Resultó tal_como la imaginaba.

Irene no terminó el postre , dejando un trozo en el plato.

Pero no fue así.

En sus labios esta investigación adquiría una alba pátina de inocencia.

Nadie en la editorial sospechó del nuevo fotógrafo.

Parecía un hombre tranquilo.

Ni_siquiera Irene supo de su vida secreta , aunque algunos indicios leves estimulaban su curiosidad.

En los meses siguientes se estrechó su relación.

El hombre se puso lentamente de pie y las invitó al interior de su morada.

Una cortina de hule aislaba un rincón del cuarto.

Mientras la madre relataba los pormenores de su desgracia , él escuchaba con los ojos entornados sumido en concentración.

Es una niña inocente.

¿ Quién puede hacerle ese perjuicio ?

Dudó del diagnóstico , pero no quiso ser descortés.

Siempre sirven para estos casos.

En esta ocasión el curandero procedió enérgicamente.

Odio ese cuarto de baño , aunque haya quedado precioso.

Baje conmigo y se convencerá.

Yo seguía sin moverme.

Cerré los ojos.

No sueles despertarte tan temprano.

Busqué su mirada pero no la encontré.

Vi que se estaba haciendo el nudo de la corbata delante del espejo.

Cambió de conversación , y en el fondo se lo agradecí.

Tal_vez fuera mejor.

Eduardo se enfadó.

Ya ves tú.

Eduardo se despidió.

Él no iba a tener tiempo de venir a buscarme.

Es horrible , su religión se lo impide.

El champán sin motivo no sabe a nada , ni_siquiera es dorado.

No pienso atender a ningún recado , llame quien llame.

No sé si conoces Nueva_York.

Por los resultados , creo que acerté.

Por_qué el oro fino perdió su brillo ? Estábamos en el bar , pedimos unos pinchos de tortilla.

Nunca lo he entendido.

Yo no tiré la toalla , me agarré a ella en una reacción incluso demasiado compulsiva, ésa es la verdad.

sin_embargo , mi trayectoria profesional , valga lo que valga , arranca de aquel enfrentamiento primero con la calamidad , de eso tampoco cabe duda.

Pero no te di facilidades para_que nos viéramos.

Cuando te colgué estuve llorando mucho rato.

Tú llevabas un vestido rojo que nunca te había visto.

En esto entraste tú en la cocina.

Te quedaste parada y nos miramos.

Lo habías oído.

Así_que lo dejaré como está.

Trataremos , más_bien , de enderezarlo.

La levanta solemnemente.

Te acuerdas de cuánto nos gustaba el mes de mayo ?

En_cambio , tu capacidad de respuesta sigue siendo asombrosa.

Fue cosa de segundos.

Los pacientes del segundo grupo son los más duros de pelar.

Se llama Raimundo.

Ahora no quiero hablar más de él.

En la exposición de Gregorio no estaba.

Nunca te han interesado los chismes.

Son los que más aspavientos hacen , pero no importan para el argumento.

Pero tú no estabas de acuerdo.