Capítulo 3

Introducción al Procesamiento del Lenguaje Natural

3.1. El Procesamiento del Lenguaje Natural

El lenguaje es uno de los aspectos fundamentales no sólo del comportamiento humano, sino de su propia naturaleza. En su forma escrita nos permite guardar un registro del conocimiento que se transmite de generación en generación, y en su forma hablada constituye el principal medio de comunicación en nuestro día a día.

El *Procesamiento del Lenguaje Natural* (NLP, *Natural Language Processing*) es la rama de las ciencias computacionales encargada del diseño e implementación de los elementos software y hardware necesarios para el tratamiento computacional del *lenguaje natural*, entendiendo como tal todo lenguaje humano, en contraposición a los *lenguajes formales* [146] propios del ámbito lógico, matemático, o computacional [110]. El objetivo último que se persigue es el de la comprensión del lenguaje humano por parte de la computadora. La consecución de un objetivo tan ambicioso, del que todavía se está muy lejos, supondría una auténtica revolución. Por una parte, los ordenadores podrían tener por fin acceso al conocimiento humano, y por otra, una nueva generación de interfaces, en lenguaje natural, facilitaría en grado sumo la accesibilidad a sistemas complejos.

3.1.1. Niveles de Análisis

Para cumplir su objetivo, un sistema de NLP necesitará hacer uso de una cantidad considerable de conocimiento acerca de la estructura del lenguaje. Este conocimiento se puede estructurar en niveles:

1. **Conocimiento morfológico**[1]**:** para determinar cómo son las palabras que constituyen el lenguaje y cómo éstas se forman a partir de unidades más pequeñas denominadas *morfemas*.

2. **Conocimiento sintáctico:** para determinar cómo se combinan las palabras para dar lugar a *sintagmas* y *frases*, así como el papel estructural que desempeña cada palabra y cada sintagma en la frase resultante.

3. **Conocimiento semántico:** para determinar el *significado* de cada palabra y cómo se construye el significado de una frase a partir de los significados de las palabras que la constituyen.

[1] También denominado conocimiento *léxico*.

4. **Conocimiento pragmático:** para determinar cómo se relaciona el lenguaje con los contextos en los que se usa.

Paralelamente a estos niveles de conocimiento se establecen cuatro niveles de análisis en los que se incluyen los diversos modelos computacionales y algoritmos para su tratamiento:

1. **Análisis morfológico**[2]**:** mediante el cual se determinan las palabras que integran un texto, así como su etiqueta morfosintáctica, utilizando para ello modelos computacionales de la morfología, basados generalmente en autómatas de estado finito, expresiones regulares, traductores de estado finito, modelos de Markov ocultos y n-gramas.

2. **Análisis sintáctico:** que realiza el agrupamiento de las palabras en sintagmas y frases mediante modelos computacionales como son las gramáticas independientes del contexto, las gramáticas lexicalizadas y las estructuras de rasgos.

3. **Análisis semántico:** mediante el cual se determina el significado de las frases de acuerdo con el significado de los sintagmas, palabras y morfemas que las forman, utilizando para ello modelos computacionales tales como la lógica de predicados de primer orden y las redes semánticas.

4. **Análisis pragmático:** que establece la identidad de las personas y objetos que aparecen en los textos, determina la estructura del discurso y gestiona el diálogo en un entorno conversacional.

En el caso del tratamiento del habla, existiría además un nivel previo de *reconocimiento del habla* y posiblemente un nivel posterior de *síntesis del habla*, los cuales harían uso de conocimiento fonético y fonológico.

3.1.2. Ambigüedad

A la hora de procesar un texto en lenguaje natural, el problema principal con el que nos hemos de enfrentar en los diferentes niveles de análisis es el de la *ambigüedad*.

A nivel morfológico, nos encontramos con que una palabra puede recibir diversas etiquetas. Por ejemplo, la palabra *sobre* puede ser un sustantivo masculino singular, una preposición, o la primera o tercera persona del presente de subjuntivo del verbo *sobrar*. En ciertos contextos la tarea de determinar la etiqueta correcta puede ser relativamente fácil, pero en frases como *"pon lo que sobre sobre el sobre"* la complejidad de este proceso es patente.

A nivel sintáctico, el hecho de que una frase sea ambigua se traduce en que es posible asociar dos o más estructuras sintagmáticas correctas a dicha frase. Tomemos el ejemplo clásico de la frase *"Juan vio a un hombre con un telescopio en una colina"*. Diferentes ubicaciones de las subestructuras correspondientes a los fragmentos *"con un telescopio"* y *"en una colina"* dan lugar a diferentes estructuras sintagmáticas de la frase, todas ellas correctas, y que se corresponden con los significados siguientes:

- Juan vio a un hombre que estaba en una colina y que tenía un telescopio;
- Juan estaba en una colina, desde donde vio a un hombre que tenía un telescopio;
- Juan estaba en una colina, desde donde miraba con un telescopio, a través del cual vio a un hombre.

[2] O análisis *léxico*.

A nivel semántico, nos encontramos con que una palabra puede tener diferentes significados o sentidos. Por ejemplo, la palabra *banda* puede referirse a:

- un grupo de personas;
- una tira de tela;
- los laterales de un barco;
- un conjunto de frecuencias del espectro radioeléctrico.

Como el significado de una frase se construye a partir de las aportaciones semánticas realizadas por las palabras que la componen, es preciso determinar en primer lugar el significado correcto de cada una de ellas. Sin embargo, el significado de una frase puede ser ambiguo incluso aun cuando las palabras que lo componen no lo son. Por ejemplo, la frase *"todos los alumnos de la facultad hablan dos idiomas"* admite dos interpretaciones distintas:

- Existen dos idiomas L y L' tales que todos los alumnos de la facultad los hablan.
- Cada uno de los alumnos de la facultad habla un par de idiomas, pero dos estudiantes distintos pueden hablar idiomas distintos.

A su vez las ambigüedades pueden ser locales o globales. Una *ambigüedad local* es aquella que surge en un momento del análisis pero que es eliminada posteriormente al analizar una porción mayor del texto. Una *ambigüedad global* es aquella que permanece una vez terminado de analizar todo el texto.

Llegados a este punto es interesante destacar que los distinto niveles de análisis no tienen porqué ser totalmente independientes entre sí, ya que, por ejemplo, y tal como hemos visto, el análisis léxico puede ofrecer diferentes etiquetas para una palabra dada, dejando que sean el analizador sintáctico e incluso el semántico los encargados de determinar aquélla más conveniente.

3.1.3. Dos Clases de Aproximaciones: Simbólica y Estadística

Es posible distinguir dos grandes tipos de aproximaciones a la hora de enfrentarse al problema del Procesamiento del Lenguaje Natural: aquéllas de carácter *simbólico*, y aquéllas de tipo empírico o *estadístico*. Hoy en día, sin embargo, parece claro que una aproximación híbrida es la más adecuada.

Aproximaciones Simbólicas

Desde sus inicios en los años 50, el Procesamiento del Lenguaje Natural ha sido abordado mediante diferentes técnicas de carácter simbólico basadas en el empleo de reglas —u otras formas de representación similares— que codifican explícitamente nuestro conocimiento del dominio, y que han sido desarrolladas por expertos humanos en el ámbito de aplicación [58, 110]. Se trata, pues, de aproximaciones *basadas en el conocimiento*, próximas a los modelos tradicionales de Inteligencia Artificial, y que precisan de una fase previa de estudio y análisis del dominio para que, de este modo, los expertos puedan identificar y describir mediante reglas las regularidades del mismo. Desde un punto de vista metodológico, se trata de una aproximación descendente, ya que intentamos imponer sobre los textos los modelos que nosotros hemos desarrollado.

Aproximaciones Estadísticas

Durante la última década, y gracias al incremento de la potencia y velocidad de los ordenadores, han cobrado especial protagonismo las aproximaciones denominadas empíricas o estadísticas, fundamentadas en el análisis y descripción estadística del lenguaje a partir de grandes corpus de texto [141, 110]. Se opta, en este caso, por un punto de vista cuantitativo, donde las diferentes posibilidades fruto de la ambigüedad lingüística son evaluadas en función de sus probabilidades asociadas empleando técnicas estadísticas. Al contrario que antes, nos encontramos ante aproximaciones ascendentes, ya que el modelo es desarrollado partiendo de los propios textos. Para ello se precisa de textos de entrenamiento sobre los que aplicar técnicas de tipo estadístico para la identificación de los patrones y asociaciones presentes en los mismos, siendo capaces incluso de capturar, en ocasiones, aspectos implícitos en el modelo que el experto es incapaz de ver.

3.2. Nivel Morfológico

En este y subsiguientes apartados abordaremos en mayor detalle los diferentes niveles de procesamiento lingüístico.

Todo lenguaje humano, sea hablado o escrito, se compone de palabras. De este modo podemos considerar a las palabras como los "ladrillos" del lenguaje. Es lógico, por tanto, empezar nuestro análisis por el procesamiento de las palabras que forman un texto. De este modo, abordaremos en nuestro primer punto el *nivel morfológico*, también referido en ocasiones como *nivel léxico*.

La *morfología* es la parte de la gramática que se ocupa del estudio de la estructura de las palabras y de sus mecanismos de formación. Las palabras están formadas por unidades mínimas de significado denominadas *morfemas* [135], los cuales podemos clasificar en dos clases: morfemas léxicos y morfemas gramaticales.

Los *morfemas léxicos*, comúnmente denominados *lexemas* o *raíces*, son los elementos que aportan el significado principal a la palabra (p.ej., *hablar*). Por el contrario, los *morfemas gramaticales*, comúnmente denominados *afijos* o, por extensión, simplemente *morfemas*, poseen únicamente significado gramatical, y nos permiten modificar el significado básico del lexema (p.ej., *hablases*).

Conforme a su posición, los afijos se clasifican en *prefijos*, antepuestos al lexema (p.ej., *innecesario*), *sufijos*, postpuestos al lexema (p.ej., *hablador*), e *infijos*, elementos que aparecen intercalados en el interior de la estructura de una palabra (p.ej., *humareda*). Desde el punto de vista de cómo éstos alteran el significado del lexema, los afijos se clasifican en flexivos y derivativos. Los *afijos flexivos* representan conceptos gramaticales tales como género y número (p.ej., *habladoras*), persona, modo, tiempo y aspecto (p.ej., *hablases*). Los *afijos derivativos*, por su parte, producen un cambio semántico respecto al lexema base, y frecuentemente también un cambio de categoría sintáctica (p.ej., *hablador*).

A la hora de estudiar las técnicas y herramientas desarrolladas a nivel morfológico en el área del Procesamiento del Lenguaje Natural nos centraremos en dos aspectos: el análisis morfológico, y la etiquetación.

3.2.1. Análisis Morfológico

El análisis morfológico de una palabra consiste en que, dada una forma de una palabra, obtener los diferentes rasgos morfológicos asociados a la misma [224], tales como su categoría gramatical, género, número, persona, etc. Por ejemplo, dada la palabra *gatos*, un analizador morfológico nos indicaría que se trata de una forma nominal masculina plural.

El análisis morfológico se encuentra íntimamente ligado a la denominada *morfología de dos niveles* [129], que considera las palabras como una correspondencia entre el *nivel léxico*, que representa la concatenación de los morfemas que constituyen una palabra, y el *nivel superficial*, que representa la forma escrita real de una palabra. De esta forma, el análisis morfológico de una palabra se lleva a cabo mediante un conjunto de reglas que hacen corresponder secuencias de letras del nivel superficial a secuencias de morfemas y rasgos morfológicos del nivel léxico. Por ejemplo, la forma superficial *gatos* se convertiría en la forma léxica *gat +Sust +Masc +Sing* mediante la cual se indica que dicha palabra es un sustantivo masculino singular.

Para realizar la correspondencia entre los niveles superficial y léxico se necesita disponer de una información mínima [121]:

1. Un **lexicón** que recoja las raíces y afijos a emplear, junto con la información básica acerca de los mismos. Por ejemplo, si se trata de una raíz nominal, verbal, etc.

2. Un modelo de ordenación para la aplicación de los morfemas, y que se conoce como **morfotácticas**. Por ejemplo, los morfemas flexivos de número se postponen al sustantivo.

3. Una serie de **reglas ortográficas** que modelen los cambios que se producen en la palabra durante la adjunción de los morfemas. Por ejemplo, en inglés, un sustantivo terminado en consonante seguido por *-y* cambia ésta por *-ie* al concatenar el morfema flexivo plural *-s*, como en el caso de *city/cities* (ciudad/ciudades).

A la hora de la implementación de esta correspondencia se utilizan traductores de estado finito [121] que se encargan de traducir un conjunto de símbolos en otro. Para esta tarea de análisis los traductores son utilizados habitualmente en cascada: primero se utiliza un traductor que reconoce el morfema léxico de las palabras y lo convierte en su forma regular, al tiempo que indica su categoría gramatical; posteriormente, se aplican traductores especializados en el reconocimiento de morfemas específicos de género, número, tiempo, persona, etc., que son transformados en rasgos morfológicos. La potencia de los traductores de estado finito viene determinada por el hecho de que la misma cascada, con las mismas secuencias de estados, puede ser utilizada tanto para obtener la forma léxica a partir de la forma superficial como para generar la forma superficial a partir de la forma léxica.

3.2.2. Etiquetación

Los problemas surgen cuando, dado un texto a analizar, nos encontramos con ambigüedades morfológicas en el mismo. Un analizador morfológico únicamente conoce la forma de la palabra, por lo que no cuenta con información suficiente para analizar correctamente cada palabra en caso de ambigüedad, ya que para ello es necesario acceder al contexto de la palabra. En una frase como *"pon lo que sobre sobre el sobre"* únicamente nos podría indicar que existen tres opciones posibles para cada aparición de la palabra *"sobre"*: sustantivo, preposición y verbo.

Al proceso de desambiguación en función del cual a cada palabra del texto le es asignado su análisis morfológico correcto —codificado por medio de una *etiqueta* (*tag*)— se le denomina *etiquetación* (*tagging*) [39], y constituye el primer paso de cara a la realización de análisis más profundos del texto, bien de carácter sintáctico o semántico. Las herramientas que implementan este proceso se denominan *etiquetadores* (*taggers*).

Fuentes de Información Relevantes para la Etiquetación

A la hora de decidir cuál es la etiqueta correcta de una palabra existen, esencialmente, dos fuentes de información [141]:

1. La primera de ellas consiste en examinar su contexto, es decir, las etiquetas de las palabras circundantes. Aunque esas palabras podrían ser también ambiguas, el hecho de observar secuencias de varias etiquetas nos puede dar una idea de cuáles son comunes y cuáles no lo son. Por ejemplo, en inglés, una secuencia como artículo-adjetivo-sustantivo es muy común, mientras que otras secuencias como artículo-adjetivo-verbo resultan muy poco frecuentes o prácticamente imposibles. Por tanto, si hubiera que elegir entre sustantivo o verbo para etiquetar la palabra `play` en la frase `a new play`, obviamente optaríamos por sustantivo.

   Este tipo de estructuras constituyen la fuente de información más directa para el proceso de etiquetación, aunque por sí misma no resulte demasiado exitosa: uno de los primeros etiquetadores basado en reglas deterministas que utilizaba este tipo de patrones sintagmáticos etiquetaba correctamente sólo el 77 % de las palabras [90]. Una de las razones de este rendimiento tan bajo es que en inglés las palabras que pueden tener varias etiquetas son muy numerosas, debido sobre todo a procesos productivos como el que permite a casi todos los sustantivos que podamos tener en el diccionario transformarse y funcionar como verbos, con la consiguiente pérdida de la información restrictiva que es necesaria para el proceso de etiquetación.

2. La segunda fuente de información consiste en el simple conocimiento de la palabra concreta, que puede proporcionarnos datos muy valiosos acerca de la etiqueta correcta. Por ejemplo, existen palabras que, aunque puedan ser usadas como verbos, su aparición es mucho más probable cuando funcionan como sustantivos. La utilidad de esta información fue demostrada de manera concluyente por Charniak, quien puso de manifiesto que un etiquetador que simplemente asigne la etiqueta más común a cada palabra puede alcanzar un índice de acierto del 90 % [52].

   La información léxica de las palabras resulta tan útil porque la distribución de uso de una palabra a lo largo de todas sus posibles etiquetas suele ser rara. Incluso las palabras con un gran número de etiquetas aparecen típicamente con un único uso o etiqueta particular.

Consecuentemente, la distribución de uso de las palabras proporciona una información adicional de gran valor, y es por ello por lo que parece lógico esperar que las aproximaciones estadísticas al proceso de etiquetación den mejores resultados que las aproximaciones basadas en reglas deterministas. En éstas últimas, uno sólo puede decir que una palabra puede o no puede ser un verbo, por ejemplo, existiendo la tentación de desechar la posibilidad de que sea un verbo cuando ésta es muy rara, creyendo que esto aumentará el rendimiento global. Por el contrario, en una aproximación estadística se puede decir *a priori* que una palabra tiene una alta probabilidad de ser un sustantivo, pero también que existe una posibilidad, por remota que sea, de ser un verbo o incluso cualquier otra etiqueta. A día de hoy, los etiquetadores modernos utilizan de alguna manera una combinación de la información sintagmática proporcionada por las secuencias de etiquetas y de la información léxica proporcionada por las palabras.

Rendimiento y Precisión de los Etiquetadores

Las cifras de rendimiento conocidas para los etiquetadores se encuentran casi siempre dentro del rango del 95 al 97 % de acierto[3]. Sin embargo, es importante señalar que estas cifras no son tan buenas como parecen, ya que implica que, en frases largas —caso de artículos periodísticos, por ejemplo—, un rendimiento del 95 % todavía supone que pueden aparecer entre una y dos palabras mal etiquetadas en cada frase. Además, estos errores no siempre se localizan en las categorías

[3]Habiéndose calculado sobre el conjunto de todas las palabras del texto. Algunos autores proporcionan la precisión sólo para los términos ambiguos, en cuyo caso las cifras serán menores.

más pobladas, tales como sustantivos, adjetivos o verbos, donde en principio parece más probable el encontrarse con palabras desconocidas, sino que muchas veces los errores aparecen asociados a las partículas que conectan los sintagmas entre sí, tales como preposiciones, conjunciones o relativos, con lo que pueden hacer que una frase tome un significado muy distinto del original.

Dejando ya de lado estas cuestiones, el rendimiento depende considerablemente de una serie de factores [141]:

- El tamaño del corpus de entrenamiento disponible. En general, a mayor disponibilidad de textos de entrenamiento, mayor y mejor será el conocimiento extraído y mejor será la etiquetación.
- El *juego de etiquetas* (*tag set*). Normalmente, cuanto más grande es el conjunto de etiquetas considerado, mayor será la ambigüedad potencial, con lo que se agrava el problema de la dispersión de datos, y la tarea de etiquetación se vuelve más compleja.
- La diferencia entre, por un lado, el diccionario y el corpus de entrenamiento empleados, y por otro, el corpus de aplicación. Si los textos de entrenamiento y los textos que posteriormente se van a etiquetar proceden de la misma fuente —por ejemplo, textos de la misma época o estilo—, entonces la precisión obtenida será mayor. Sin embargo, si los textos de aplicación pertenecen a un periodo o género distintos —p.ej., textos científicos contra textos periodísticos—, entonces el rendimiento será menor.
- Las palabras desconocidas. Un caso especial del punto anterior es la cobertura del diccionario. La aparición de palabras desconocidas puede degradar el rendimiento, situación común, por ejemplo, al intentar etiquetar material procedente de algún dominio técnico.

Un cambio en cualquiera de estas cuatro condiciones puede producir un fuerte impacto en la precisión alcanzada por el etiquetador. Es importante señalar que estos factores son externos al proceso de etiquetación y al método elegido para realizar dicho proceso, siendo su efecto a menudo mucho mayor que la influencia ejercida por el propio método en sí.

Etiquetación Basada en Reglas

Los primeros etiquetadores abordaban el problema de la desambiguación mediante aproximaciones basadas en reglas empleando una arquitectura en dos etapas [100, 128]. En una primera fase se le asigna a cada palabra una lista de sus etiquetas potenciales en base a un diccionario. Es entonces cuando, en una segunda etapa, se aplican las reglas de desambiguación para identificar la etiqueta correcta.

El primer algoritmo para la asignación de etiquetas que se conoce estaba incorporado en el analizador sintáctico utilizado en el proyecto TDAP, implementado entre 1958 y 1969 en la Universidad de Pennsylvania [100]. Anteriormente, los sistemas de procesamiento del lenguaje natural utilizaban diccionarios con información morfológica de las palabras pero, que se sepa, no realizaban desambiguación de etiquetas. El sistema TDAP realizaba esta desambiguación mediante 14 reglas escritas a mano que eran ejecutadas en un orden basado en la frecuencia relativa de las etiquetas de cada palabra.

Poco después del TDAP surgió el sistema CGC de Klein y Simmons [128], con sus tres componentes: un lexicón, un analizador morfológico y un desambiguador por contexto. El pequeño diccionario de 1.500 palabras incluía aquellas palabras raras que no podían ser tratadas por el analizador morfológico, tales como sustantivos, adjetivos y verbos irregulares. El analizador morfológico utilizaba los sufijos flexivos y derivativos para asignar un conjunto de etiquetas a cada palabra. En ese momento entraban en acción un conjunto de 500 reglas

encargadas de seleccionar la etiqueta correcta, consultando para ello las islas de palabras contiguas no ambiguas. El juego de etiquetas constaba de 30 etiquetas.

Etiquetación Estocástica

Actualmente, uno de los modelos de etiquetación más extendidos, es el de la utilización de procedimientos estadísticos basados en la probabilidad de aparición conjunta de secuencias de n palabras o *n-gramas*. La matemática subyacente a los n-gramas fue propuesta por primera vez por Markov [143], quien utilizó bigramas y trigramas para predecir si la siguiente letra de una palabra rusa sería una vocal o una consonante. Shannon [216] aplicó posteriormente los n-gramas para calcular aproximaciones a las secuencias de palabras en inglés. A partir de los años 50, y gracias al trabajo de Shannon, los modelos de Markov fueron ampliamente utilizados para modelar secuencias de palabras. En décadas posteriores su uso decayó, principalmente debido a la argumentación de muchos lingüistas, entre ellos Chomsky [53], de que los modelos de Markov eran incapaces de modelar completamente el conocimiento gramatical humano. Los modelos de n-gramas resurgen en los años 70 al hacerse públicos los trabajos realizados en el centro de investigación Thomas J. Watson de IBM [115, 27] y en la Universidad de Carnegie Mellon [29], en los que se utilizan con éxito n-gramas para tareas de reconocimiento del habla.

En los años 70 se creó el corpus Lancaster-Oslo/Bergen (LOB) de inglés británico. Para su etiquetación se utilizó el etiquetador CLAWS [145], basado en un algoritmo probabilístico que puede considerarse una aproximación al enfoque actual basado en la utilización de modelos de Markov ocultos. El algoritmo utilizaba la probabilidad de aparición conjunta de dos etiquetas, pero en lugar de almacenar dicha probabilidad directamente, la clasificaba como *rara* ($P(\text{etiqueta} \mid \text{palabra}) < 0,01$), *infrecuente* ($0,01 \leq P(\text{etiqueta} \mid \text{palabra}) < 0,10$) o *normalmante frecuente* ($P(\text{etiqueta} \mid \text{palabra}) \geq 0,10$).

El etiquetador probabilístico de Church [55] seguía una aproximación muy cercana a la de los modelos de Markov ocultos, extendiendo la idea de CLAWS para asignar la probabilidad real a cada combinación palabra/etiqueta, utilizando el algoritmo de Viterbi [259, 75] para encontrar la mejor secuencia de etiquetas. Sin embargo, al igual que CLAWS, almacenaba la probabilidad de una etiqueta dada la palabra para calcular

$$P(\text{etiqueta} \mid \text{palabra}) \times P(\text{etiqueta} \mid n \text{ etiquetas anteriores})$$

en lugar de almacenar la probabilidad de una palabra dada la etiqueta, tal y como actualmente hacen los etiquetadores basados en modelos de Markov ocultos para calcular

$$P(\text{palabra} \mid \text{etiqueta}) \times P(\text{etiqueta} \mid n \text{ etiquetas anteriores})$$

Los etiquetadores posteriores ya introdujeron explícitamente la utilización de modelos de Markov ocultos. Tal es el caso del etiquetador TnT de Brants [37], y MrTagoo de Graña [83] que constituyen claros ejemplos de las herramientas recientes de alto rendimiento que utilizan modelos de Markov ocultos basados en n-gramas.

Antes de describir en qué consiste un *modelo de Markov oculto*, debemos describir en qué consiste un *modelo de Markov observable* [141]. Consideremos un sistema que en cada instante de tiempo se encuentra en un determinado estado. Dicho estado pertenece a un conjunto finito de estados Q. Regularmente, transcurrido un espacio de tiempo discreto, el sistema cambia de estado de acuerdo con un conjunto de probabilidades de transición asociadas a cada uno de los estados del modelo. Los instantes de tiempo asociados a cada cambio de estado se denotan como $t = 1, 2, \ldots, T$, y el estado actual en el instante de tiempo t se denota como q_t. En general, una descripción probabilística completa del sistema requeriría la especificación del estado actual,

así como de todos los estados precedentes. Sin embargo, las cadenas de Markov presentan dos características de suma importancia:

1. La *propiedad del horizonte limitado*, que permite truncar la dependencia probabilística del estado actual y considerar, no todos los estados precedentes, sino únicamente un subconjunto finito de ellos. Una cadena de Markov de orden n es la que utiliza n estados previos para predecir el siguiente estado. Por ejemplo, para el caso de las cadenas de Markov de tiempo discreto de primer orden tenemos que $P(q_t = j|q_{t-1} = i, q_{t-2} = k, \dots) = P(q_t = j|q_{t-1} = i)$, es decir, dependería únicamente del estado anterior; en caso de ser de segundo orden, de los dos estados anteriores, y así sucesivamente.

2. La *propiedad del tiempo estacionario*, que nos permite considerar sólo aquellos procesos en los cuales $P(q_t = j|q_{t-1} = i)$ es independiente del tiempo, lo que a su vez nos lleva a definir una matriz de probabilidades de transición independientes del tiempo $A = \{a_{ij}\}$, donde $\forall i, j; \quad 1 \leq i, j \leq N; \quad a_{ij} = P(q_t = j|q_{t-1} = i) = P(j|i)$ y se cumplen las restricciones estocásticas estándar: $a_{ij} \geq 0$ para todo i y j, y $\sum_{j=1}^{N} a_{ij} = 1$ para todo i. Adicionalmente, es necesario especificar el vector $\pi = \{\pi_i\}$ que almacena la probabilidad $\pi_i \geq 0$ que tiene cada uno de los estados de ser el estado inicial: $\forall i; 1 \leq i \leq N; \pi_i = P(q_1 = i)$.

A un proceso estocástico que satisface estas características se le puede llamar un *modelo de Markov observable*, porque su salida es el conjunto de estados por los que pasa en cada instante de tiempo, y cada uno de estos estados se corresponde con un suceso observable. Esta modelización puede resultar demasiado restrictiva a la hora de ser aplicada a problemas reales. A continuación extenderemos el concepto de modelos de Markov de tal manera que sea posible incluir aquellos casos en los cuales la observación es una función probabilística del estado. El modelo resultante, denominado *modelo de Markov oculto* (HMM, *Hidden Markov Model*), es un modelo doblemente estocástico, ya que uno de los procesos no se puede observar directamente (está oculto), y sólo se puede observar a través de otro conjunto de procesos estocásticos, los cuales producen la secuencia de observaciones. Un HMM se caracteriza por la 5-tupla (Q, V, π, A, B) donde:

1. $Q = \{1, 2, \dots, N\}$ es el conjunto de estados del modelo. Aunque los estados permanecen ocultos, para la mayoría de las aplicaciones prácticas se conocen a priori. Por ejemplo, para el caso de la etiquetación de palabras, cada etiqueta del juego de etiquetas utilizado sería un estado. Generalmente los estados están conectados de tal manera que cualquiera de ellos se puede alcanzar desde cualquier otro en un solo paso, aunque existen muchas otras posibilidades de interconexión. El estado actual en el instante de tiempo t se denota como q_t. El uso de instantes de tiempo es apropiado, por ejemplo, en la aplicación de los HMM al procesamiento de voz. No obstante, para el caso de la etiquetación de palabras, no hablaremos de los instantes de tiempo, sino de las posiciones de cada palabra dentro de la frase.

2. V es el conjunto de los distintos sucesos que se pueden observar en cada uno de los estados. Por tanto, cada uno de los símbolos individuales que un estado puede emitir se denota como $\{v_1, v_2, \dots, v_M\}$. En el caso de la etiquetación de palabras, M es el tamaño del diccionario y cada $v_k, 1 \leq k \leq M$, es una palabra distinta.

3. $\pi = \{\pi_i\}$, es la distribución de probabilidad del estado inicial, cumpliéndose que $\pi_i \geq 0$, $\forall i; \quad 1 \leq i \leq N; \quad \pi_i = P(q_1 = i)$, y $\sum_{i=1}^{N} \pi_i = 1$.

4. $A = \{a_{ij}\}$ es la distribución de probabilidad de las transiciones entre estados, esto es, $\forall i, j, t; \quad 1 \leq i \leq N, 1 \leq i \leq N, 1 \leq t \leq T; \quad a_{ij} = P(q_t = j|q_{t-1} = i) = P(j|i)$, cumpliéndose que $a_i, j \geq 0$ y que $\sum_{j=1}^{N} a_{ij} = 1$ para todo i.

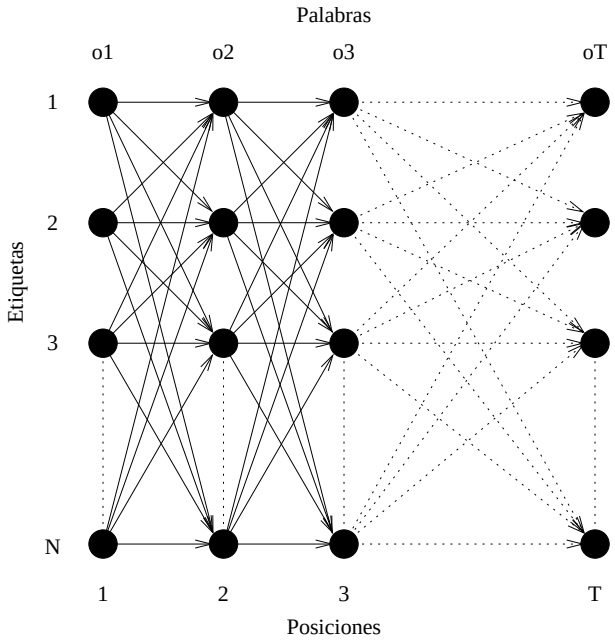


Figura 3.1: Enrejado genérico de T observaciones y N estados

5. $B = \{b_j(v_k)\}$ es la distribución de probabilidad de los sucesos observables, es decir, $\forall j, k, t; \quad 1 \leq j \leq N, 1 \leq k \leq M, 1 \leq t \leq T; \quad b_j(v_k) = P(o_t = v_k | q_t = j) = P(v_k|j)$, cumpliéndose que $\sum_{k=1}^{M} b_j(v_k) = 1$ para todo j. Este conjunto de probabilidades se conoce también con el nombre de *conjunto de probabilidades de emisión*.

Los parámetros del modelo —las probabilidades de transición y las probabilidades de salida de los estados— son estimados mediante un proceso de entrenamiento a partir de un corpus previamente desambiguado manualmente a tal efecto [37].

En base a dicho modelo, y dada una secuencia de observaciones (palabras) $O = (o_1, o_2, \ldots, o_T)$, $o_i \in V$, queremos determinar la secuencia de estados $S = (q_1, q_2, \ldots, q_T)$ óptima, es decir, aquélla que mejor *explica* la secuencia de observaciones. De una forma más sencilla, dada una secuencia de palabras O a etiquetar, queremos determinar la secuencia de etiquetas S más probable. Para ello se genera el *enrejado* o *diagrama de Trellis* correspondiente a dicha secuencia y modelo, tal como se aprecia en la figura 3.1, y que recoge todas las secuencias posibles de etiquetas para dicho secuencia de palabras. Sobre este enrejado se calculará la secuencia de etiquetas más probable empleando el algoritmo de Viterbi [259, 75]. De hecho, en el caso concreto de la etiquetación de palabras, los cálculos involucrados en el algoritmo de Viterbi se realizan frase por frase sobre enrejados simplificados como el de la figura 3.2, donde en cada posición no se consideran todos los estados posibles —o sea, todas la etiquetas del juego de etiquetas utilizado—, sino sólo las etiquetas candidatas que proponga el diccionario para cada palabra.

Etiquetación Basada en Transformaciones

Algunas de las hipótesis de funcionamiento de los modelos de Markov no se adaptan bien a las propiedades sintácticas de los lenguajes naturales, por lo que surge inmediatamente la idea de utilizar modelos más sofisticados que puedan establecer condiciones no sólo sobre las etiquetas precendentes, sino también sobre las palabras precedentes, o que permitan emplear contextos

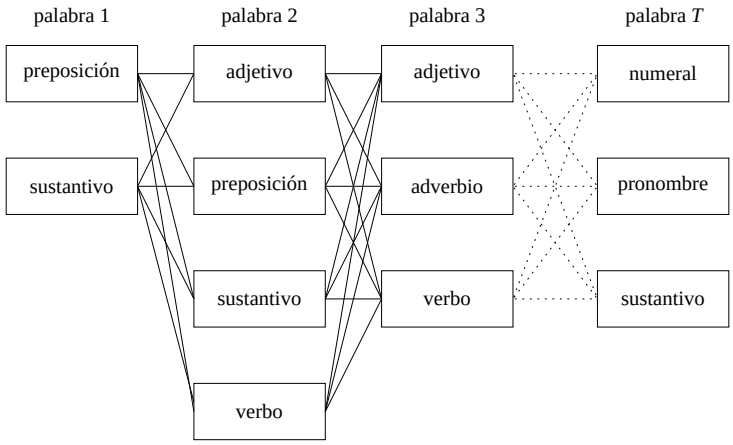


Figura 3.2: Enrejado simplificado para la etiquetación de una frase de T palabras

mayores a los asumibles empleando modelos de Markov[4]. Bajo estas premisas, Brill definió un sistema de etiquetación basado en reglas [38] que a partir de un corpus de entrenamiento infiere automáticamente las reglas de transformación. El así denominado *etiquetador de Brill* alcanza una corrección comparable a la de los etiquetadores estocásticos y, a diferencia de éstos, la información lingüística no se captura de manera indirecta a través de grandes tablas de probabilidades, sino que se codifica directamente bajo la forma de un pequeño conjunto de reglas no estocásticas muy simples, pero capaces de representar interdependencias muy complejas entre palabras y etiquetas.

El proceso de etiquetación consta de tres partes, que se infieren automáticamente a partir de un corpus de entrenamiento: un etiquetador léxico, un etiquetador de palabras desconocidas, y un etiquetador contextual:

1. Un *etiquetador léxico*, que etiqueta inicialmente cada palabra con su etiqueta más probable, sin tener en cuenta el contexto en el que dicha palabra aparece. Esta etiqueta se estima previamente mediante el estudio del corpus de entrenamiento. A las palabras desconocidas se les asigna en un primer momento la etiqueta correspondiente a sustantivo propio si la primera letra es mayúscula, o la correspondiente a sustantivo común en otro caso. Posteriormente, el etiquetador de palabras desconocidas aplica en orden una serie de reglas de transformación léxicas. Si se dispone de un diccionario previamente construido, es posible utilizarlo junto con el que el etiquetador de Brill genera automáticamente.

2. Un *etiquetador de palabras desconocidas*, que opera justo después de que el etiquetador léxico haya etiquetado todas las palabras presentes en el diccionario, y justo antes de que se apliquen las reglas contextuales. Este módulo intenta *adivinar* una etiqueta para una palabra desconocida en función de su sufijo, de su prefijo, y de otras propiedades relevantes similares. Básicamente, cada transformación consta de dos partes: una descripción del contexto de aplicación, y una regla de reescritura que reemplaza una etiqueta por otra.

3. Un *etiquetador contextual*, que actúa justo después del etiquetador de palabras desconocidas, aplicando en orden una secuencia de reglas contextuales que, al igual que las léxicas, también han sido previamente inferidas de manera automática a partir del corpus de entrenamiento.

[4]El orden de los HMM está limitado a valores pequeños debido a la carga computacional que implican y a la gran cantidad de nuevos parámetros que necesitaríamos estimar.

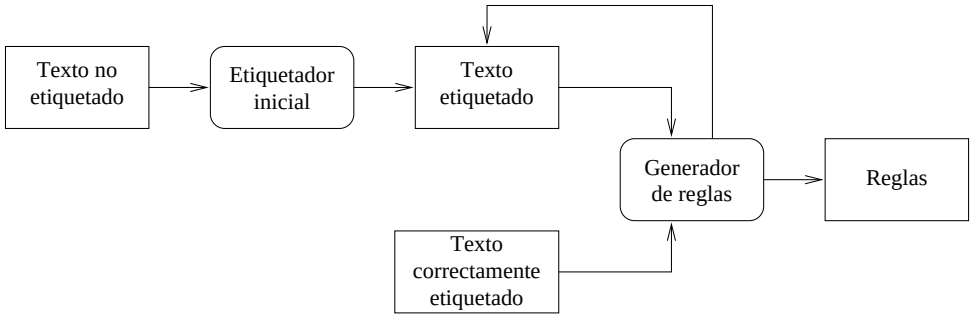


Figura 3.3: Proceso de aprendizaje de reglas en un etiquetador de Brill

El proceso de aprendizaje de las reglas, tanto las léxicas en el caso del etiquetador de palabras desconocidas, como las contextuales en el caso del etiquetador contextual, selecciona el mejor conjunto de transformaciones y determina su orden de aplicación. El algoritmo consta de los pasos que se ilustran en la figura 3.3. En primer lugar, se toma una porción de texto no etiquetado, se pasa a través de la fase o fases de etiquetación anteriores, se compara la salida con el texto correctamente etiquetado, y se genera una lista de errores de etiquetación con sus correspondientes contadores. Entonces, para cada error, se determina qué instancia concreta de la plantilla genérica de reglas produce la mayor reducción de errores. Se aplica la regla, se calcula el nuevo conjunto de errores producidos, y se repite el proceso hasta que la reducción de errores cae por debajo de un umbral dado.

La técnica de etiquetación de Brill resulta considerablemente más lenta que las basadas en modelos probabilísticos. No sólo el proceso de entrenamiento consume una gran cantidad de tiempo, sino que el proceso de etiquetación es también inherentemente lento. La principal razón de esta ineficiencia computacional es la potencial interacción entre las reglas, de manera que el algoritmo puede producir cálculos innecesarios.

Etiquetación Basada en Gramáticas de Restricciones

Las técnicas para la etiquetación de textos vistas hasta ahora son las que podríamos denominar *clásicas*. No obstante, estos métodos difícilmente permiten sobrepasar la cota del 96 % de precisión obtenida. Por otra parte, en el caso de los etiquetadores estocásticos esta cifra se reduce todavía más cuando los corpus de entrenamiento y aplicación son de tipos distintos.

Estas deficiencias abrieron paso a investigaciones sobre nuevos métodos de etiquetación, fruto de las cuales es el paradigma de *etiquetación mediante reglas de restricción*. Dentro de este campo, el sistema de etiquetación por excelencia es el sistema EngCG[5] [264]. En este sistema encontramos un conjunto de reglas escritas a mano que manejan el contexto global o, mayormente, el contexto local. No existe una verdadera noción de gramática formal, sino más bien una serie de restricciones, casi siempre negativas, que van eliminando sucesivamente los análisis imposibles de acuerdo con el contexto [207]. La idea es similar al aprendizaje basado en transformaciones, excepto por el hecho de que es un humano, y no un algoritmo, el que modifica iterativamente el conjunto de reglas de etiquetación para minimizar el número de errores. En cada iteración, el conjunto de reglas se aplica al corpus y posteriormente se intentan modificar dichas reglas de manera que los errores más importantes queden manualmente corregidos.

Podría pensarse que se trata de un retroceso a los métodos tradicionales basados en reglas, sin embargo la idea general en la que se basa este nuevo planteamiento consiste en la utilización de reglas de menor compromiso para evitar así errores en situaciones dudosas. De este modo se

[5] *English Constraint Grammar.*

ha logrado obtener una serie de métodos de alta precisión, con el inconveniente de que en algunas palabras la ambigüedad no ha sido eliminada por completo después del proceso de etiquetación, ya que no utiliza reglas de compromiso máximo. A pesar de esto, la mayoría de las palabras tendrán una única etiqueta tras el proceso de etiquetación.

Por otra parte, existe también la posibilidad de emplear este formalismo en combinación con un etiquetador tradicional como, por ejemplo, un etiquetador estocástico, que sería el encargado de completar el proceso de desambiguación. Esta solución, estudiada por el autor de esta memoria en [85], consiste en podar el enrejado inicial mediante la aplicación de reglas de restricción, eliminando combinaciones de etiquetas imposibles. Sobre el enrejado resultante se aplicaría el algoritmo de Viterbi para proceder a la desambiguación final.

El empleo de este nuevo paradigma basado en restricciones parece ofrecer mejores resultados que los etiquetadores basados en modelos de Markov ocultos —en torno al 99 % en el caso del sistema EngCG—, especialmente cuando los corpus de entrenamiento y de aplicación no provienen de la misma fuente, ya que las reglas son, en principio, universales, al no haber sido extraídas a partir de un corpus de entrenamiento. Sin embargo, la comparación de estos dos modelos es difícil de realizar, ya que cuando el sistema EngCG no es capaz de resolver determinadas ambigüedades, éste devuelve el conjunto de etiquetas obtenido para la palabra. El problema de esta técnica es, al igual que en los modelos tradicionales basados en reglas, la necesidad de participación de expertos lingüistas para la creación de las reglas, lo que supone un problema en comparación con el aprendizaje automático de los HMMs.

La Real Academia Española está desarrollando también un formalismo de reglas de restricciones denominado sistema Rtag [223]. Este sistema aplica gramáticas de reglas de contexto ponderadas sobre textos anotados ambiguamente. De esta forma, cuando un contexto satisface la descripción estructural de una regla, recibe la puntuación que indica la regla. Esta puntuación puede ser positiva, para promover lecturas, o negativa, para penalizarlas. Una vez finalizado el proceso, permanecen las lecturas con mayor puntuación siempre que estén por encima de un umbral definido previamente. El sistema también intenta eliminar lecturas imposibles en función del contexto, sin pérdida de lecturas posibles aunque éstas sean poco probables. Para la poda de lecturas en función del contexto se utiliza información derivada del propio texto (características estructurales, tipográficas o secuenciales), información gramatical (sobre todo concordancia y restricciones de aparición conjunta) e información gramatical estructural (toma de decisiones con ayuda de la información estructural derivable de la secuencia lineal del texto).

Otros Paradigmas de Etiquetación

Existen también otros paradigmas de etiquetación a mayores de los descritos anteriormente, algunos de los cuales presentaremos brevemente.

Ratnaparkhi emplea *modelos de máxima entropía* en su etiquetador JMX [181]. Esta técnica, de naturaleza también probabilística, combina las ventajas de los etiquetadores basados en transformaciones y de los etiquetadores estocásticos basados en modelos de Markov, ya que se trata de una técnica de gran flexibilidad que permite manejar un abanico de propiedades del lenguaje mayor que los modelos de Markov, acercándose al caso de Brill, y que además, al generar las distribuciones de probabilidad de etiquetas para cada palabra, permite su integración dentro de un marco probabilístico.

Los *árboles de decisión* son también empleados en tareas de etiquetación, como en el caso del etiquetador TreeTagger [215]. Un *árbol de decisión* se puede ver como un mecanismo que etiqueta todas las hojas dominadas por un nodo con la etiqueta de la clase mayoritaria de ese nodo. Posteriormente, a medida que descendemos por el árbol, reetiquetamos las hojas de los nodos hijos, si es que difieren de la etiqueta del nodo padre, en función de las respuestas

a las cuestiones o decisiones que aparecen en cada nodo. Esta manera de ver los árboles de decisión guarda ciertas similitudes con el aprendizaje basado en transformaciones, ya que ambos paradigmas realizan series de reetiquetados trabajando con subconjuntos de datos cada vez más pequeños.

Otro de los paradigmas clásicos de computación, las *redes de neuronas artificiales*, es también empleado en tareas de etiquetación. Este es el caso de la propuesta de Marques y Lopes [144] para el portugués.

Queda patente, pues, el amplio abanico de posibilidades a la hora de implementar un etiquetador gracias a la continua investigación sobre el tema. Muestra de ello es, por ejemplo, el reciente desarrollo de aproximaciones basadas en *algoritmos evolutivos* [25] o *support vector machines* [81].

3.3. Nivel Sintáctico

Una vez identificadas y analizadas las palabras individuales que componen un texto, el siguiente paso lógico consiste en estudiar cómo éstas se organizan y relacionan entre sí para formar unidades superiores (sintagmas y frases), y las funciones que representan las unidades inferiores dentro de la unidad superior. Se trata, por lo tanto, de estudiar la estructura sintáctica del texto.

3.3.1. Conceptos Básicos: Lenguajes, Gramáticas y Ambigüedad

La acotación de un lenguaje, la obtención de una representación manejable del mismo, es un paso necesario para posibilitar su procesamiento. La forma más simple de lograr este objetivo es enumerar sus cadenas constituyentes, pero este procedimiento resulta poco práctico cuando el lenguaje consta de más de unas pocas cadenas o pretendemos definir propiedades o clasificaciones entre los lenguajes. De ahí que surja la necesidad de establecer algún mecanismo para generar lenguajes con una notación finita. Estos generadores de lenguajes son las *gramáticas*, sistemas matemáticos adaptados al tratamiento computacional. De este modo definimos una *gramática* como una 4-tupla $\mathcal{G} = (N, \Sigma, P, S)$ donde:

- Σ es el alfabeto finito de la gramática o conjunto finito de *símbolos terminales*, o *palabras*, o *categorías léxicas*,
- N es un conjunto finito de *símbolos no terminales*, o *variables*, o *categorías sintácticas*, $N \cap \Sigma = \emptyset$,
- P es un subconjunto finito de $(N \cup \Sigma)^* N (N \cup \Sigma)^* \times (N \cup \Sigma)^*$ a cuyos elementos denominaremos *producciones*, *reglas*, o *reglas de producción*, y
- $S \in N$ es el *símbolo inicial*, o *axioma* de la gramática.

Con frecuencia se prefiere representar las producciones $(\alpha, \beta) \in P$ como $\alpha \to \beta \in P$. Al primer miembro α de una regla de producción $\alpha \to \beta$ se le suele llamar *parte izquierda* de la regla de producción, mientras que el segundo miembro β recibe el nombre de *parte derecha* de la regla. A las reglas cuya parte derecha es la cadena vacía ε, reglas de la forma $\alpha \longrightarrow \varepsilon$, se les llama *reglas-$\varepsilon$* o *producciones-$\varepsilon$*. Cuando dos producciones $\alpha \to \beta$ y $\alpha \to \gamma$ tienen la misma parte izquierda, se pueden escribir abreviadamente como $\alpha \to \beta \mid \gamma$.

De esta forma, un ejemplo de gramática sería aquélla que genera el lenguaje los numeros binarios pares, es decir, aquéllos terminados en 0:

$$\mathcal{G} = (\{S\}, \{0, 1\}, \{S \to A0, A \to 0A, A \to 1A, A \to \varepsilon\}, S) \tag{3.1}$$

Las cadenas del lenguaje se construyen partiendo del símbolo inicial S, siendo las producciones las encargadas de describir cómo se lleva a cabo esa generación. Empleando las reglas de producción de la gramática, se pueden construir distintas secuencias de símbolos terminales y no terminales a partir del símbolo inicial. Se denominará *formas sentenciales* a dichas secuencias, que podemos definir recursivamente de la siguiente manera. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, entonces:

- S es una *forma sentencial*.
- Si $\alpha\beta\gamma$ es una forma sentencial y $\beta \to \delta \in P$, entonces $\alpha\delta\gamma$ también es una *forma sentencial*.

Intuitivamente, S es la forma sentencial más simple. A partir de ella se generan las demás formas sentenciales. Dada una forma sentencial y una regla de producción se generará una nueva forma sentencial sustituyendo una aparición de la parte izquierda de la regla en la primera, por la parte derecha de dicha regla. Un tipo especialmente interesante de forma sentencial es aquella que está formada exclusivamente por símbolos terminales. De esta forma, dada una gramática $\mathcal{G} = (N, \Sigma, P, S)$, denominaremos *frase generada por una gramática* a cualquier forma sentencial que únicamente contenga símbolos terminales. Las frases son, por lo tanto, cadenas de símbolos terminales obtenidas a través de la aplicación de reglas de producción de la gramática[6], partiendo del símbolo raíz S. Por lo tanto, son las cadenas que formarán parte del lenguaje generado por la gramática.

A modo de ejemplo, y retomando de nuevo la gramática definida en 3.1 para la generación de binarios pares, tenemos que:

Siendo S forma sentencial,	dado que $S \to A0 \in P$,	$A0$ es forma sentencial.
Siendo $A0$ forma sentencial,	dado que $A \to 0A \in P$,	$0A0$ es forma sentencial.
Siendo $0A0$ forma sentencial,	dado que $A \to 1A \in P$,	$01A0$ es forma sentencial.
Siendo $01A0$ forma sentencial,	dado que $A \to \varepsilon \in P$,	010 es una frase.

La generación de formas sentenciales y frases descrita anteriormente puede formalizarse empleando el concepto de *derivación*. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, se define una *derivación directa* o *derivación en un solo paso*, $\Rightarrow$, como sigue:

$$\text{Si } \alpha\beta\gamma \in (N \cup \Sigma)^* \text{ y } \beta \to \delta \in P, \text{ entonces } \alpha\beta\gamma \Rightarrow \alpha\delta\gamma.$$

En el caso de una cadena de derivaciones directas, se dirá que $\alpha\beta\gamma$ *deriva indirectamente* $\alpha\delta\gamma$ si y sólo si:

- $\beta \Rightarrow \delta_1 \Rightarrow \delta_2 \ldots \Rightarrow \delta_n \Rightarrow \delta$, que notaremos $\alpha\beta\gamma \overset{+}{\Rightarrow} \alpha\delta\gamma$, o bien
- $\beta = \delta$ ó $\alpha\beta\gamma \overset{+}{\Rightarrow} \alpha\delta\gamma$, que notaremos $\alpha\beta\gamma \overset{*}{\Rightarrow} \alpha\delta\gamma$

En caso de conocer el número exacto k de derivaciones directas, se usará la notación $\alpha\beta\gamma \overset{k}{\Rightarrow} \alpha\delta\gamma$.

Por otra parte, la gramática impone una estructura arborescente sobre la frase o forma sentencial generada, de tal modo que dada una regla $\alpha \to \beta$, ésta conforma en sí misma un árbol donde el nodo raíz es el símbolo de la parte izquierda, siendo sus nodos hijo los símbolos de la parte derecha. Esta estructura arborescente se denomina *arbol sintáctico* o *de derivación* [182]. A modo de ejemplo, y continuando el ejemplo de los números binarios pares, recogemos en la figura 3.4 el árbol sintáctico correspondiente al número 010.

Las formas sentenciales, frases incluídas, serán aquellas que se pueden derivar a partir del símbolo inicial de la gramática. El conjunto de todas las frases generadas por una gramática

[6] Las reglas de producción que hemos usado para generar unas formas sentenciales a partir de otras.

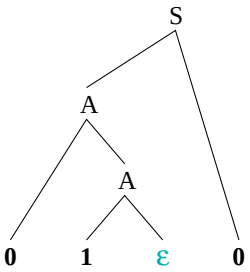


Figura 3.4: Árbol sintáctico del número binario 010

forma un lenguaje sobre el alfabeto Σ de la gramática, que podemos definir formalmente de la siguiente manera. Sea $\mathcal{G} = (N, \Sigma, P, S)$ una gramática, el *lenguaje generado por la gramática* es el conjunto $L(\mathcal{G})$ definido del siguiente modo:

$$L(\mathcal{G}) = \left\{ w | w \in \Sigma^*,\ S \overset{*}{\Rightarrow} w \right\}$$

Finalmente, introduciremos el concepto de *ambigüedad*, que se produce cuando para una misma forma sentencial existe más de un árbol sintáctico válido. En base a ello podemos definir los conceptos de gramática y lenguaje ambiguos, de tal forma que se dice que una gramática $\mathcal{G} = (N, \Sigma, P, S)$ es una *gramática ambigua* si y sólo si $\exists x \in L(\mathcal{G})$, para la cual existen al menos dos árboles sintácticos válidos. Asimismo, diremos que un lenguaje L *no es ambiguo* si y sólo si existe una gramática $\mathcal{G}$ no ambigua tal que $L(\mathcal{G}) = L$. En caso contrario diremos que L es un *lenguaje ambiguo*.

Tomemos como ejemplo una pequeña gramática aproximativa de las oraciones *sujeto-verbo-complemento* con reglas

$$\begin{aligned}
&\text{S} \rightarrow \text{NP}\ \ \text{VP}\\
&\text{S} \rightarrow \text{S}\ \ \text{PP}\\
&\text{NP} \rightarrow \text{Sust}\\
&\text{NP} \rightarrow \text{Det}\ \ \text{Sust}\\
&\text{NP} \rightarrow \text{NP}\ \ \text{PP}\\
&\text{PP} \rightarrow \text{Prep}\ \ \text{NP}\\
&\text{VP} \rightarrow \text{Verbo}\ \ \text{NP}
\end{aligned}$$

Esta gramática resulta ambigua puesto que la frase *"Juan vio un hombre con un telescopio"* puede ser generada de dos formas diferentes, dando lugar a dos árboles sintácticos distintos, tal y como se aprecia, en línea contínua y discontínua, en la figura 3.5.

3.3.2. Jerarquía de Chomsky

Dependiendo de la forma de las reglas de producción, podremos obtener lenguajes más o menos complejos. De este modo, podemos clasificar los lenguajes en función de las gramáticas que los generan y, más concretamente, en función de la forma de dichas reglas de producción. Así, Chomsky [54] propone una jerarquía con cuatro clases. En ella se clasifican, de menor a mayor complejidad, las gramáticas formales y sus lenguajes asociados, de forma que cada nivel de la jerarquía incluye a las gramáticas y lenguajes del nivel anterior, tal como se muestra en la figura 3.6.

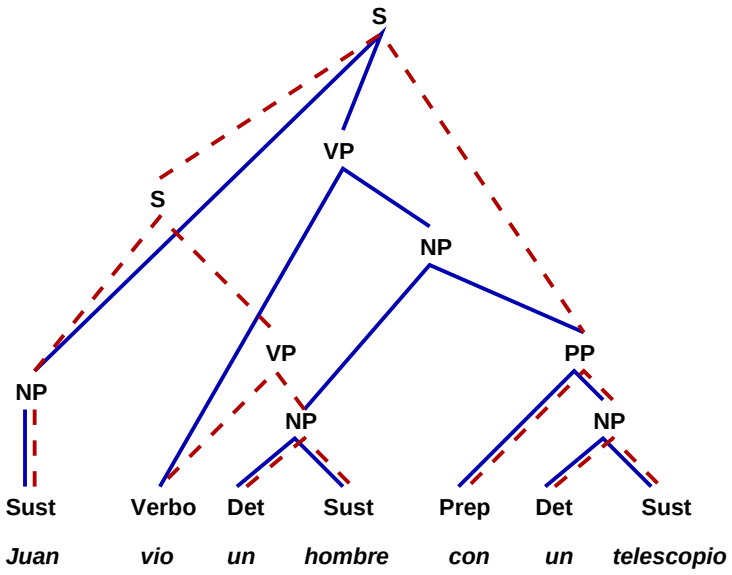


Figura 3.5: Ejemplo de ambigüedad sintáctica

Gramáticas regulares. En este caso, las producciones son de la forma: $A \rightarrow x$ ó $A \rightarrow xB$. Este tipo de producciones nos asegura que todas las formas sentenciales generadas contendrán a lo sumo un único símbolo no terminal. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes regulares*.

Gramáticas independientes del contexto. Sus producciones tienen un único símbolo no terminal en la parte izquierda: $A \rightarrow \beta$. De esta forma, a la hora de realizar un paso de derivación directo, es posible decidir qué símbolo no terminal queremos reescribir independientemente del contexto que lo rodea. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes independientes del contexto*.

Gramáticas dependientes del contexto. La parte izquierda de las producciones pueden contener cualquier combinación de símbolos terminales y no terminales, siempre y cuando sea de longitud menor o igual que la parte derecha. De esta forma aseguramos que al aplicar una derivación sobre una forma sentencial obtendremos otra forma sentencial de igual o mayor longitud. Las producciones siguen el patrón $\alpha \rightarrow \beta$, $|\alpha| \leq |\beta|$, siendo $|\alpha|$ la longitud de α, esto es, el número de símbolos en α. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes sensibles al contexto*.

Gramáticas con estructura de frase. No existe ninguna restricción sobre las producciones. Los lenguajes que pueden ser generados por este tipo de gramáticas se denominan *lenguajes recursivamente enumerables*.

En el caso de los *lenguajes naturales*, no se sabe a ciencia cierta qué lugar ocuparían en esta jerarquía, aunque se cree que estarían situadas entre los lenguajes independientes del contexto y los lenguajes dependientes del contexto, posiblemente más cerca de los primeros que de los segundos, tal y como podemos apreciar en la figura 3.6. Esta suposición se basa en el hecho de que la mayoría de las construcciones sintácticas sólo dependen *suavemente* del contexto en el cual son aplicadas.

Debemos reseñar que la jerarquía de Chomsky no es la única forma de clasificar lenguajes (por ejemplo, las *gramáticas contextuales* [142] son ortogonales a la jerarquía de Chomsky), aunque sí la más común.

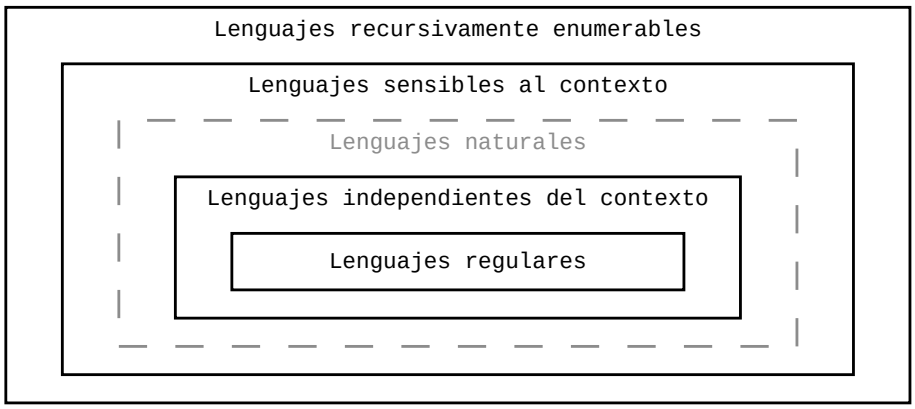


Figura 3.6: Diagrama de Venn correspondiente a la jerarquía de Chomsky

3.3.3. Análisis Sintáctico

Hasta ahora nos hemos centrado en dos conceptos fundamentales, el de lenguaje como un conjunto de cadenas y el de gramática como formalismo descriptivo de un lenguaje. El problema del análisis sintáctico se centra en encontrar un mecanismo que sirva para establecer la gramaticalidad de una cadena, es decir, reconocer si ésta pertenece al lenguaje generado por la gramática, y proponer una representación apropiada de dicho proceso de análisis. Los algoritmos que realizan sólo la primera de las dos acciones se denominan *reconocedores sintácticos*, mientras que a aquéllos capaces de generar además una representación del proceso —es decir, capaces de obtener el árbol sintáctico de la cadena procesada— se les denomina *analizadores sintácticos*. En este punto, podemos introducir una primera clasificación de los algoritmos de análisis sintáctico:

- Los *algoritmos ascendentes* son aquellos que construyen el árbol desde las hojas hasta la raíz.
- Los *algoritmos descendentes* actúan en sentido contrario a los ascendentes, de la raíz a las hojas.
- Las *estrategias mixtas* combinan los dos enfoques anteriores. Aunque existen algoritmos puros, tanto ascendentes como descendentes, lo más habitual es hacer uso de estas estrategias, que de alguna forma aportan lo mejor de cada mundo.

Podemos igualmente establecer clasificaciones de algoritmos de análisis sintáctico basándonos en otros criterios. El primero de éstos es el tratamiento del posible no determinismo en el análisis, factor de especial importancia en el caso de los lenguajes naturales debido a su ambigüedad inherente:

- *Algoritmos basados en retroceso.* En estos algoritmos el no determinismo se simula mediante un mecanismo de retroceso [13]. Cuando varias alternativas son posibles, se escoge sólo una, y, si ésta resulta infructuosa, se retrocede hasta el último punto de no determinismo y se escoge otra. Los cálculos realizados en las alternativas exploradas anteriormente se desechan. Este enfoque es sencillo, pues economiza espacio y recursos, pero presenta varios problemas:
  - Los cálculos realizados en las alternativas exploradas anteriormente se desechan. Por tanto, si éstos vuelven a ser son necesarios en una alterntiva posterior, deberán ser calculados de nuevo.

  - El criterio de selección de las alternativas puede no ser óptimo, llevándonos a una elección incorrecta de alternativas que no conducen a una solución y, por tanto, a cálculos innecesarios.
  - En caso de ambigüedad de la gramática, puede haber varias soluciones diferentes. Si se desea encontrarlas todas, se deberá forzar el retroceso tanto si se encuentran soluciones como si no, agravando los problemas anteriores.

- *Algoritmos basados en programación dinámica.* Mediante técnicas de programación dinámica [46, 65, 67], se almacenan los cálculos ya realizados de forma que no sea necesario repetirlos en caso de que se vuelvan a necesitar. Esto nos permite, incluso, compartir cálculos entre las diversas alternativas de análisis derivadas de una gramática ambigua, solucionando en parte los problemas de los algoritmos basados en retroceso, en particular la multiplicación innecesaria de cálculos y los problemas de no terminación.

Otra posible clasificación de los algoritmos de análisis sintáctico es en función de su dependencia de la estructura gramatical durante el análisis:

- *Guiados por la gramática.* La elección de las alternativas se realiza con la información proporcionada por las reglas de producción.

- *Guiados por control finito.* En estos algoritmos existe una fase de pre-procesamiento antes del análisis. En ella, se utiliza la información de las reglas de la gramática para construir un mecanismo de control que se encargará de la elección de alternativas durante el proceso de análisis.

En el contexto del lenguaje natural, ambiguo, complejo, y propenso a contener errores, cobran protagonismo, frente a las técnicas clásicas de *análisis sintáctico completo* o convencional, ciertos tipos de análisis sintáctico capaces de abordar esta problemática:

- *Análisis sintáctico robusto.* Al contrario que ocurre con los lenguajes formales, en el lenguaje natural no siempre es posible conseguir una cadena de entrada correcta y completa —debido, por ejemplo, al uso incorrecto de la lengua por parte del interlocutor—, ni una gramática exhaustiva que cubra todas las posibles cadenas de entrada —debido a su complejidad. Esta situación nos obliga a realizar el análisis sintáctico en presencia de lagunas gramaticales e, incluso, de errores. A este tipo de análisis se le califica de *robusto* [246, 245]. Debemos precisar que esta clase de análisis está dirigido a obtener la mayor cantidad de información posible a partir de una cadena de entrada con errores. Otra aproximación diferente sería intentar corregir dichos errores para obtener un análisis sintáctico completo [60]. Ambas soluciones no son, sin embargo, excluyentes, pudiendo combinarse [247, 248].

- *Análisis sintáctico parcial.* Emplearemos este término para referirnos a las técnicas de análisis capaces no sólo de obtener, de ser posible, el análisis completo de una entrada, sino también, en su defecto, sus posibles subanálisis [197, 198, 257, 47].

- *Análisis sintáctico superficial.* No siempre es necesario realizar un análisis detallado de la estructura sintáctica del texto. Para algunas tareas basta realizar un análisis *superficial* de la misma [94, 92], identificando únicamente las estructuras de mayor entidad, tales como frases nominales, grupos preposicionales, etc. En este contexto es común la utilización de cascadas de autómatas o traductores finitos [11, 10].

3.3.4. Formalismos Gramaticales

Existen diferentes *formalismos gramaticales* que pueden ser empleados a la hora de abordar el problema del análisis sintáctico en lenguaje natural.

A partir de los años 60, la mayor parte de los modelos computacionales para el procesamiento del lenguaje natural se basaron en gramáticas independientes del contexto debido a la disponibilidad de algoritmos eficientes para realizar el análisis de este tipo de gramáticas, tales como el CYK [271, 123] o el algoritmo de Earley [67].[7]

También es frecuente extender las gramáticas independientes del contexto mediante la decoración de producciones y árboles de análisis con probabilidades para así posibilitar un mejor tratamiento de las ambigüedades [36]. De cara a su análisis se desarrollaron extensiones análogas de los correspondientes algoritmos clásicos de análisis [116, 228].

Sin embargo, las lenguas naturales presentan construcciones que no pueden ser descritas mediante gramáticas independientes del contexto. Surge entonces la necesidad de contar con formalismos más adecuados que permitan llenar el hueco descriptivo existente.

Una de las posibilidades es la del empleo de la operación de unificación en entornos gramaticales [125, 56]. Entre los formalismos con unificación más extendidos se encuentran las *gramáticas de cláusulas definidas*, una generalización de las gramáticas independientes del contexto basada en lógica de primer orden [171]. Sobre la base de una gramática independiente del contexto, se generalizan los símbolos de la misma añadiendo información adicional, *atributos* del símbolo. De este modo los símbolos de la gramática nos permiten representar un conjunto infinito de elementos, extendiendo de este modo su dominio de definición. A continuación se establece una operación que nos permita la manipulación de los símbolos gramaticales con atributos y se adapta convenientemente el mecanismo de derivación de la gramática de forma que tenga en cuenta la información contenida en éstos. La extensión se realiza mediante términos lógicos de primer orden, considerando la unificación [195] como mecanismo de manipulación.

Otros formalismos que utilizan unificación, en este caso unificación de estructuras de rasgos, son las gramáticas léxico-funcionales [122, 169], las gramáticas con estructura de frase dirigidas por el núcleo [178], y las gramáticas categoriales de unificación [234].

Puesto que la estructura sintáctica asociada a las frases es una estructura jerárquica representada normalmente como un árbol o, en el caso de frases ambiguas, como un conjunto de árboles, parece natural pensar que un formalismo que manipule árboles y que presente cierta dependencia suave del contexto resultaría adecuado para la descripción de los fenómenos sintácticos que aparecen en el lenguaje natural. Con este objetivo nacen las *gramáticas de adjunción de árboles* [119], uno de los formalismos gramaticales derivados de las gramáticas independientes del contexto más ampliamente difundidos. En este tipo de gramáticas la estructura fundamental es el árbol, en lugar de la producción. Los árboles se clasifican en iniciales y auxiliares. Los árboles iniciales suelen utilizarse para representar las estructuras de las frases elementales, mientras que los árboles auxiliares se utilizan para representar estructuras recursivas mínimas que se pueden añadir a otros árboles. Los árboles se combinan mediante las operaciones de adjunción y sustitución. Desde el punto de vista lingüístico las grandes ventajas de las gramáticas de adjunción de árboles provienen de su carácter lexicalizado —ya que permiten asociar una palabra con cada árbol— y de su dominio de localidad extendido, posibilitando el establecimiento de relaciones de larga distancia entre los nodos de árboles elementales. También en este caso existen adaptaciones de los algoritmos clásicos de análisis para el caso de las gramáticas de adjunción de árboles [213]. Debemos destacar también la investigación se ha hecho en torno al análisis sintáctico de gramáticas de adjunción de árboles, tanto en análisis

[7] Una visión conjunta de la mayor parte de los algoritmos de análisis sintáctico para gramáticas independientes del contexto puede encontrarse en la obra de Sikkel [217].

bidireccional [20, 16], como mediante autómatas [16, 66].

Existen multitud de formalismos equivalentes a las gramáticas de adjunción de árboles. Entre ellos destacan las gramáticas lineales de índices [18, 19], las gramáticas categoriales combinatorias [225], y las gramáticas de núcleo [186]. Todos estos formalismos se engloban en la clase de los formalismos gramaticales suavemente sensibles al contexto [120].

Existen otros formalismos gramaticales que no se basan en las gramáticas independientes del contexto. Por ejemplo, las *gramáticas de dependencia* [150], que se fundamentan en las relaciones existentes entre palabras y no en las relaciones entre constituyentes.

3.4. Nivel Semántico

La *semántica* es el estudio del significado lingüístico. Consecuentemente, a la hora de realizar un análisis semántico de un texto, nuestro objetivo será el de obtener el significado de las frases que lo componen. En este apartado realizaremos una breve introducción a este campo, menos detallada que en el caso de los niveles anteriores, ya que el nivel semántico, al igual que el nivel pragmático, no es abordado profundamente en nuestro trabajo.

El primer punto a abordar es el de las *representaciones semánticas*, ya que las diferentes aproximaciones al análisis semántico parten de la base de que la semántica de los diferentes elementos lingüísticos —palabras, sintagmas— puede ser capturada mediante estructuras formales. Estas estructuras deberían cumplir una serie de características:

1. **Verificabilidad.** Debemos ser capaces de determinar la verdad o falsedad acerca del enunciado expresado por nuestra representación de acuerdo a nuestra *base de conocimiento*.

2. **No ambigüedad.** Si bien pueden existir ambigüedades lingüísticas a nivel semántico, como en el caso de la frase *"todos los alumnos de la facultad hablan dos idiomas"*, no debemos confundir esta ambigüedad en el enunciado con una ambigüedad en la representación de dicho enunciado. Por lo tanto, independientemente de la existencia de ambigüedades en el texto fuente, el tipo de representación semántica empleada debe admitir una única interpretación no ambigua, interpretación que en su caso sí deberá reflejar la ambigüedad del enunciado.

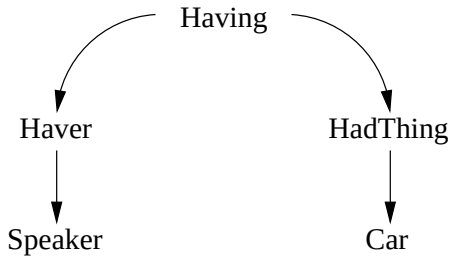
3. **Existencia de una forma canónica.** Debemos ser capaces de asociar una única representación a entradas diferentes con formas diferentes pero igual significado. De este modo evitaremos el riesgo de evaluar de diferente manera la verdad o falsedad de una aserción según la manera en que ésta hubiese sido formulada. Esto supone tratar la *variación lingüística* del lenguaje, es decir, cómo un mismo concepto puede ser expresado de formas diferentes mediante el empleo, por ejemplo, de sinónimos (p.ej., *listo/inteligente*), construcciones gramaticales equivalentes (p.ej., *Juan asesinó a Pedro/Pedro fue asesinado por Juan*), etc.

4. **Disponibilidad de mecanismos de inferencia y uso de variables.** De esta forma el sistema deberá ser capaz de decidir acerca de la verdad o falsedad de proposiciones que no estén explícitamente representadas en su base de conocimiento, pero que sí sean derivables a partir de la misma. Por su parte, el empleo de variables permitirá el manejo de entradas con referencias no totalmente definidas.

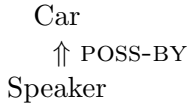
5. **Expresividad.** El tipo de representación empleada debe ser capaz de representar cualquier aserción de interés para la aplicación.

$$\exists\, x, y\ \text{Having}(x) \wedge \text{Haver}(\text{Speaker}, x) \wedge \text{HadThing}(y, x) \wedge \text{Car}(y)$$

(a) Predicado lógico de primer orden



(b) Red semántica



(c) Diagrama de dependencia conceptual

```
Having
  Haver:      Speaker
  HadThing:   Car
```

(d) *Frame*

Figura 3.7: Representaciones semánticas de la oración *"I have a car"* ("Yo tengo un coche")

La figura 3.7 recoge, para el ejemplo *"I have a car"* ("Yo tengo un coche"), algunas de estructuras formales de representación semántica más utilizadas, y comunes al ámbito de la Inteligencia Artificial clásica [185].

La primera de ellas recoge una aproximación basada en el *cálculo de predicados de primer orden*, una de las soluciones más extendidas. Los inicios de su empleo para la captura del significado de textos en lenguaje natural data de la década de los 60, cuando Woods [267] investiga la posibilidad de utilizar representaciones basadas en lógica de predicados para los sistemas de búsqueda de respuestas en lugar de representaciones ad-hoc como venía siendo corriente hasta entonces.

Por esa misma época, aquellos investigadores interesados en el modelado cognitivo del lenguaje y de la memoria trabajaban en varias formas de representación basadas en redes asociativas. Es en este periodo cuando se comienza a investigar con profusión en el ámbito de las *redes semánticas* [147], el segundo caso recogido en la figura 3.7. En una red semántica los objetos son representados como nodos en un grafo, mientras que las relaciones entre los mismos son representadas mediante arcos etiquetados.

La tercera de las estructuras es un diagrama de *dependencia conceptual* [214]. Se trata de una forma de representación de amplio uso en el campo del lenguaje natural, y que emplea una serie de primitivas conceptuales que se pueden combinar entre sí para expresar un significado dado.

El último caso recogido en la figura 3.7 se trata de una representación basada en *frames*, estructuras de conocimiento que constan de una *cabecera*, que identifica el *frame*, y de una serie

de atributos —denominados *slots*—, que pueden contener tanto valores atómicos como nuevos *frames* anidados.

A la hora de realizar el *análisis semántico* propiamente dicho —y contando ya con una estructura de representación adecuada—, nuestro objetivo es el de obtener la representación semántica de la frase componiendo de algún modo las representaciones individuales de sus componentes. Uno de los enfoques más utilizados es el denominado *análisis dirigido por la sintaxis* (*syntax-driven semantic analysis*) [121]. Éste se basa en el llamado *principio de composicionalidad*[8], y según el cual la semántica de una objeto puede ser obtenida a partir de la semántica de sus componentes. Fue Montague [166] quien mostró que el enfoque composicional podía ser aplicado a una parte importante del lenguaje natural, introduciendo la estructura de modelos teóricos en la teoría lingüística, y dando lugar de este modo a una integración mucho más fuerte entre las teorías de la sintaxis formal y un amplio rango de estructuras semánticas.

Sin embargo, si bien el significado de una frase puede obtenerse a partir de los significados de las palabras y sintagmas que la componen, también es cierto que los meros significados aislados de los mismos no son suficientes. De esta forma, si partimos de un conjunto de palabras {*Juan*, *matar*, *Pedro*}, no es en absoluto lo mismo decir *"Juan mató a Pedro"* que *"Pedro mató a Juan"*. Por lo tanto, debemos matizar nuestra afirmación anterior, ya que el significado de una frase no se obtiene únicamente a partir de las palabras que la forman, sino que también viene dado por la forma en que éstas se relacionan entre sí. En otras palabras, el significado de la frase depende parcial pero inexorablemente de su estructura sintáctica. De esta forma, en el *análisis dirigido por la sintaxis* el sistema parte de las representaciones de significado de los componentes para, guiado por la estructura o sintaxis de la frase, obtener la representación resultante de la frase.

En relación a lo anterior, debemos destacar que uno de los entornos aplicativos más representativos en los cuales se trata de capturar la semántica de los textos es el de la propia *Recuperación de Información*, puesto que, como ya se apuntó en el apartado 2.2.1, la mayor parte de los sistemas de recuperación de información actuales están basados en una interpretación extrema del principio de composicionalidad, al considerar que la semántica de los documentos reside únicamente en las palabras que lo forman, sin tener en cuenta el orden de los constituyentes ni su estructura sintáctica. Es lo que se conoce habitualmente como aproximación basada en *bag-of-terms*.

Uno de las herramientas más utilizadas en tareas de procesamiento semántico es la base de datos lexicográfica WordNet [158, 156, 97, 70, 33], en el caso del inglés, o su equivalente EuroWordNet [263], en el caso de otras lenguas europeas —ya abordadas en el apartado 2.4.1.

El hecho de que una misma palabra pueda tener diversos significados según el contexto en el que ésta se utilice constituye uno de los principales problemas del análisis semántico. Las técnicas de *desambiguación del sentido de las palabras* [226, 68] tratan de resolver esta ambigüedad léxica seleccionando el sentido adecuado de cada palabra en una frase. La complejidad de esta tarea viene determinada por la cantidad de palabras homónimas y polisémicas presentes en el vocabulario del idioma. En esencia, se aplican técnicas similares a las utilizadas para realizar la etiquetación de las palabras en el nivel morfológico, pero en lugar de utilizar etiquetas morfosintácticas se utilizan etiquetas semánticas que identifican el sentido de las palabras. Por tanto se tratará de obtener el sentido más probable de una palabra en relación con los sentidos de las palabras vecinas.

[8] Comúnmente conocido como *principio de composicionalidad de Frege*, aún cuando Frege nunca se refirió explícitamente a él.

3.5. Nivel Pragmático

La *pragmática* es el estudio de la relación entre el lenguaje y el contexto en el que se utiliza. El contexto incluye elementos como la identidad de las personas y los objetos participantes, y por tanto la pragmática incluye el estudio de cómo se utiliza el lenguaje para referenciar a personas y cosas. También incluye el contexto del discurso y, por consiguiente, el estudio de cómo se estructura el discurso y de cómo los participantes en una conversación gestionan el diálogo. En consecuencia, para realizar el análisis pragmático se precisa de algoritmos para la resolución de la anáfora, modelos computacionales para recuperar la estructura de monólogos y diálogos, y modelos de gestión del diálogo.

La importancia de la correcta interpretación de la *anáfora* viene dada por su necesidad a la hora de procesar correctamente textos escritos en lenguaje natural [159], especialmente en el caso de tareas como la extracción de información y la creación de resúmenes de textos. Los primeros trabajos sobre resolución de la anáfora trataban de explotar el conocimiento lingüístico y del dominio que se tenía, el cual era difícil tanto de representar como de procesar, requiriendo una notable participación humana. La necesidad de desarrollar soluciones robustas de bajo coste computacional hizo que muchos investigadores optasen por técnicas que hiciesen uso de un conjunto limitado de recursos lingüísticos. Este enfoque vino también motivado por la existencia de herramientas fiables y eficientes para el tratamiento de corpus, tales como etiquetadores-lematizadores y analizadores sintácticos superficiales.

En lo referente al procesamiento de *diálogos*, los primeros sistemas conversacionales, como el ELIZA [265], eran sistemas muy simples, basados fundamentalmente en el emparejamiento de patrones. Se hizo necesaria una mejor comprensión de los mecanismos del diálogo humano para el desarrollo de gestores del diálogo más sofisticados. Se estableció, por ejemplo, el concepto de subdiálogo, y se observó que los diálogos orientados a la realización de una determinada tarea presentaban una estructura cercana a la de la tarea que estaba siendo realizada. En el caso del *monólogo*, su tratamiento es similar al del diálogo, si bien menos complejo, ya que por ejemplo el tratamiento de la anáfora requiere analizar, en el diálogo, tanto el texto del actuante como el de los otros interlocutores.

En la actualidad uno de los principales ámbitos de aplicación del análisis pragmático es el de la *traducción automática* (*machine translation*) [107]. Las primeras investigaciones en este campo se remontan al década de los 50. El optimismo inicial dio paso, al poco tiempo, a una etapa de oscurantismo debido a la falta de recursos software y hardware adecuados para la tarea. Si bien algunos investigadores siguieron trabajando en el campo —caso del sistema SYSTRAN [5]— fue a partir de los 80 cuando cobró nuevo interés. Frente a las primeras aproximaciones de esta década, basadas en el significado y en la utilización de una interlingua, la investigación actual gira en torno a la utilización de métodos estadísticos y basados en la alineación de corpus multilingüe paralelos [184, 109], gracias a la disponibilidad de corpus de gran tamaño y de herramientas computacionales de suficiente potencia. Este nuevo interés radica en el aumento de las relaciones comerciales internacionales, la puesta en práctica de políticas gubernamentales que propician la traducción de documentos oficiales a varias lenguas —caso de la Unión Europea—, y la difusión mediante Internet de una ingente cantidad de información en formato electrónico.

En la misma línea, y por su relación con la temática de esta tesis, llamamos la atención sobre un campo de investigación en continuo desarrollo desde hace algunos años: la *Recuperación de Información Translingüe* (CLIR, *Cross-Lingual Information Retrieval*) [93]. Se trata de uno de los campos dentro de la Recuperación de Información, y en el cual consultas y documentos están en idiomas diferentes.

3.6. Procesamiento del Lenguaje Natural y Recuperación de Información

La comunidad científica que investiga la Recuperación de Información ha mostrado en repetidas ocasiones su interés por el empleo de técnicas de Procesamiento de Lenguaje Natural. La razón para este interés reside en el hecho de que decidir acerca de la relevancia de un documento dado respecto a una consulta consiste, en esencia, en decidir acerca de si el texto del documento satisface la necesidad de información expresada por el usuario, lo que implica que el sistema debe comprender, en cierta medida, el contenido de dicho documento [229].

Tal y como ya hemos indicado anteriormente, los sistemas de IR actuales se basan en una interpretación extrema del *principio de composicionalidad*, que nos dice que la semántica de un documento reside únicamente en los términos que lo forman [121]. De este modo, podemos suponer que cuando una palabra determinada está presente en un documento, dicho documento trata del tema indicado por dicha palabra [130]. De igual modo, cuando una consulta y un documento comparten términos índice, se puede presumir que el documento aborda, de algún modo, el tema sobre el que trata la consulta [24] (véase apartado 2.2.1). En base a ello ambos, consultas y documentos, son representados mediante conjuntos de términos índice o palabras clave —paradigma *bag-of-terms* [26]—, de tal forma que la decisión acerca de la relevancia o no de un documento respecto a una consulta es tomada de acuerdo al grado de correspondencia entre el conjunto de términos índice asociados al documento y el conjunto de términos índice asociados a la consulta. Asimismo, la utilización de pesos a la hora de medir el mayor o menor poder discriminante de un determinado término (véase apartado 2.2.2), así como el empleo de funciones de ordenación (véase apartado 2.2.3), permiten la ordenación de los documentos pertenecientes al conjunto respuesta de acuerdo a su grado de relevancia respecto a la consulta.

En este contexto, una de las principales limitaciones a las que han de hacer frente los sistemas de IR es la *variación lingüística* inherente al lenguaje humano [24], es decir, aquellas alteraciones de carácter lingüístico que un término puede sufrir y que impiden el correcto establecimiento de correspondencias —con el correspondiente detrimento de precisión y cobertura— en situaciones como la existencia de cambios en la flexión de una palabra —p.ej., *gato* vs. *gatas*—, el empleo de sinónimos —p.ej., *matar* vs. *asesinar*—, la presencia de ambigüedades semánticas —p.ej. *banda* (de tela) vs. *banda* (de forajidos)—, etc.

Se hace patente, pues, que el lenguaje no es un mero repositorio de palabras, tal como pretende el paradigma *bag-of-terms*, sino que nos permite comunicar conceptos, entidades, y relaciones, de múltiples maneras diferentes. Del mismo modo, las palabras se combinan a su vez en unidades lingüísticas de mayor complejidad, cuyo significado no siempre viene dado por el significado de sus palabras componente.

La aplicación de técnicas de Procesamiento del Lenguaje Natural al ámbito de la Recuperación de Información surge como respuesta a la necesidad de mejorar el tratamiento de la variación lingüística. El desarrollo de nuevas herramientas de NLP, más eficientes, robustas, y precisas, así como la cada vez mayor potencia de las nuevas generaciones de ordenadores han promovido el desarrollo de dicha aplicación. Sin embargo, debemos precisar a este respecto que el trabajo de investigación llevado a cabo hasta la fecha ha estado primordialmente centrado en el caso del inglés, y si bien otras lenguas como el francés o el alemán han sido también objeto de estudio, el español ha quedado relegado frecuentemente a un segundo plano. Por otra parte, la mayor complejidad lingüística del español frente al inglés en todos sus niveles no permite una extrapolación inmediata al español de los resultados obtenidos para el inglés, requiriendo la realización de experimentos específicos.

A continuación describiremos los diferentes niveles de variación lingüística existentes, así como las diferentes aproximaciones propuestas para abordar estos niveles.

3.6.1. Variación Morfológica

La *morfología* es la parte de la gramática que se ocupa del estudio de la estructura de las palabras y de sus mecanismos de formación en base a unidades mínimas de significado denominadas *morfemas* (ver apartado 3.2). Dentro de la morfología podemos hablar de morfología flexiva y morfología derivativa. La *morfología flexiva* hace referencia a aquellos cambios predecibles fruto de las variaciones de género y número (p.ej., *hablador* vs. *habladoras*), persona, modo, tiempo y aspecto (p.ej., *hablar* vs. *hablases*), etc., los cuales no conllevan una modificación de la categoría gramatical de la palabra, ni tampoco cambios relevantes de significado. Por contra, la *morfología derivativa* estudia la formación de nuevo léxico en base a mecanismos de *derivación*, la unión de morfemas individuales o grupos de morfemas —en este caso morfemas derivativos— para formar términos más complejos. Al contrario que en el caso de la flexión, las modificaciones derivativas sí producen un cambio semántico respecto al término original, y frecuentemente también un cambio de categoría sintáctica (p.ej., *hablar* vs. *hablador*).

La *variación morfológica* conlleva, por tanto, una pérdida de cobertura por parte del sistema, ya que impide establecer correspondencias entre términos próximos debido a las alteraciones morfológicas flexivas o derivativas que ha sufrido. Las soluciones clásicas a la hora de mitigar los efectos de la variación de carácter morfológico pasan por la *expansión de la consulta* mediante las variantes morfológicas de los términos originales [168], o por el empleo de técnicas de *stemming*. Ambas técnicas fueron ya introducidas en los apartados 2.4.1 y 2.3.1, respectivamente, y si bien su efecto es equivalente, la técnica más extendida a la hora de su empleo para la normalización morfológica de un texto es el *stemming*.

Sin embargo, las técnicas tradicionales de *stemming* —el algoritmo de Porter, por ejemplo—, son bastante agresivas, pudiendo dar lugar a normalizaciones erróneas que incidan negativamente en la precisión. Por ejemplo, en inglés, un algoritmo basado en Porter normalizaría las palabras *general* (general) y *generous* (generoso), en una forma común *gener-*. Este problema se agrava en el caso de lenguas de morfología más compleja e irregular que la del inglés [24, 233], como ocurre en el caso del español [74].

A nivel flexivo, Arampatzis et al. [24] proponen una solución más conservadora en la que el proceso de normalización retenga la categoría gramatical de la palabra original. Para ello se propone el empleo de técnicas de *lematización*, en las que los términos que componen el texto sean reducidos a su *lema* o forma canónica —forma masculina singular en nombres y adjetivos e infinitivo en verbos—, eliminando de esta forma la flexión de una palabra. La aproximación al nivel derivativo debe ser, sin embargo, más cauta, debido a los cambios semánticos y de categoría gramatical que conllevan con frecuencia las relaciones derivativas. Algunas relaciones podrían venir indicadas por la propia sintaxis, tales como la nominalización de la acción de un verbo, mientras que otras relaciones más indirectas podrían requerir el empleo de información semántica. No obstante, el potencial de su uso, especialmente en el caso de lenguajes de morfología rica —como el español—, es notable [209, 233, 114].

3.6.2. Variación Semántica

La *variación semántica* viene dada por la *polisemia*, el hecho de que una misma palabra pueda tener diferentes significados o sentidos en función de su contexto. Tal es el caso, por ejemplo, de *banda*: banda de música, banda de delincuentes, banda de tela, etc. Esto incide negativamente en la precisión del sistema, ya que una consulta referente a, por ejemplo, *bandas municipales* podría devolver, equivocadamente, documentos sobre *bandas de delincuentes*.

Para reducir en lo posible la variación semántica de un texto se hace preciso recurrir entonces a técnicas de *desambiguación del sentido de las palabras* [226, 68] para identificar el sentido concreto de cada palabra. Dichas técnicas fueron ya tratadas en el apartado 3.4

3.6.3. Variación Léxica

La *variación léxica* hace referencia a la posibilidad de emplear términos diferentes a la hora de representar un mismo significado, como ocurre en el caso de los *sinónimos*. Este tipo de variación lingüística incide también negativamente en la cobertura del sistema, ya que una consulta que hiciese referencia al término *automóvil* no devolvería documentos que únicamente se refiriesen al término *coche*.

A la hora de tratar estos fenómenos debe tenerse en cuenta el gran impacto que la variación semántica tiene en los procesos de tratamiento de la variación léxica, ya que la elección de uno u otro término semánticamente equivalente a una palabra dada depende del sentido de la misma en su contexto. Es por ello que a la hora de tratar la variación léxica se hace necesario eliminar, en primer lugar, la variación semántica del texto mediante procesos de desambiguación del sentido. Se estima, de hecho, que una desambiguación con una efectividad menor del 90 % puede ser incluso contraproducente [208] en este tipo de procesos, si bien otros trabajos, como el de Stokoe et al. [227] apuntan a que una efectividad del 50 %-60 % es suficiente.

Algunas de las soluciones propuestas para este problema pasan por la expansión de consultas con términos relacionados léxico-semánticamente —sinónimos, hipónimos, etc.—, el empleo de distancias conceptuales a la hora de comparar consultas y documentos, y la indexación mediante *synsets* de WordNet [158, 156, 97, 70, 33]. Asimismo, es precisamente esta base de datos léxica, WordNet, la fuente de información semántica más común.

La expansión de consultas mediante términos relacionados léxico-semánticamente ha sido empleada en repetidas ocasiones, mostrando buenos resultados en el caso de consultas cortas o incompletas, pero escasa o nula incidencia en el caso de consultas suficientemente completas [261].

Por otra parte, experimentos empleando recuperación basada en distancias semánticas [222] han mostrado mejoras en los resultados, si bien dichos experimentos fueron limitados, por lo que no pueden considerarse plenamente representativos.

Finalmente, la indexación mediante *synsets* [82] en lugar de palabras únicamente produce mejoras cuando el sentido de las palabras de las consultas ha sido plenamente desambiguado.

3.6.4. Variación Sintáctica

El tratamiento de la *variación sintáctica*, fruto de las modificaciones en la estructura sintáctica de un discurso manteniendo su significado, han sido tratadas tradicionalmente mediante dos aproximaciones diferentes: aquéllas que operan sobre estructuras sintácticas, y aquéllas que emplean frases a modo de *términos índice complejos*. En ambos casos el objetivo perseguido es aumentar la precisión en el proceso de recuperación, salvando en lo posible las limitaciones del paradigma *bag-of-terms* [233] a la hora de considerar la información sintáctica del texto.

El empleo de representaciones complejas en base a estructuras sintácticas durante el proceso de indexación y/o búsqueda, como podrían ser el caso de árboles [182, 256] o grafos [167], plantea problemas debido a su alto coste, haciéndolas poco adecuadas para su empleo a gran escala en entornos prácticos.

La solución más extendida pasa por el empleo de frases como términos índice dentro de un paradigma de recuperación clásico. La hipótesis sobre la que se sustenta su uso es la de que las frases denotan conceptos o entidades más significativos que en el caso de las palabras individuales, por lo que presumiblemente deberían constituir términos índice más precisos y descriptivos [230, 24]. En lo que respecta a la cobertura del sistema, ésta no se ve inicialmente afectada, ya que los términos simples que componen de una frase hubieran también dado lugar a correspondencias entre documento y consulta de haber empleado únicamente términos simples [161].

Tradicionalmente se han considerado dos tipos de frases en IR: las frases *estadísticas*, obtenidas mediante técnicas estadísticas que buscan secuencias de palabras contiguas que coocurren con una frecuencia significativa [162, 42], y las frases *sintácticas*, formadas por conjuntos de palabras relacionadas sintácticamente, y obtenidas mediante técnicas de NLP [168, 130, 112, 172, 106]. La mayor utilidad de uno u otro tipo de frases en tareas de IR es una cuestión todavía por discernir plenamente, aunque existen resultados que apuntan hacia las frases sintácticas como mejor opción, al menos en un futuro a medio plazo ante la presumible disponibilidad de técnicas de análisis y desambiguación sintáctica adecuadas [24]. Por otra parte, debemos puntualizar que gran parte de las soluciones investigadas hasta ahora en el caso de las soluciones sintácticas suelen emplear como términos índice complejos únicamente sintagmas nominales [132, 161, 106]. Es también común, tanto en el caso de frases estadísticas como sintácticas, que los términos complejos empleados consten nada más que de dos constituyentes, descomponiendo de ser preciso aquellos términos de más de dos constituyentes en compuestos de únicamente dos elementos [24, 172, 69].

Debe tenerse también en cuenta que los términos complejos son utilizados mayormente en combinación con términos simples [168, 161, 106, 230, 42], ya que el empleo único de frases como términos índice permite capturar sólamente una vista parcial e insuficiente del documento, lo que redunda en un empeoramiento de los resultados [161].