Universidad de los Andes
Facultad de Ingeniería
Postgrado en Computación

Propuesta de Tesis:

Procesamiento del Lenguaje Natural basado en una "gramática de estilos" para el idioma español

Autor: Hilda Yelitza Contreras Z.
Tutor: Prof. Jacinto Dávila

Marzo, 2001

Resumen: Procesamiento del lenguaje natural basado en una gramática de estilos para el idioma español.

Este proyecto pretende desarrollar una herramienta para interpretar documentos en español y extraer de ellos buenos descriptores. Los problemas de procesar el lenguaje natural y de extraer información, han sido atacados desde hace varías décadas [Moreno, A.:1998], [Allen, J.:1995], [Wilson, B.:1998]. Sin embargo, las investigaciones no han sido suficientes para diseñar un sistema que rinda como un humano al interpretar lenguaje natural. El lenguaje natural escapa a todos los esfuerzos de tratamiento computacional, al parecer, debido a que el conocimiento lingüístico está asociado, de formas sutiles y desconocidas, con el conocimiento contextual que tiene el hablante. En este trabajo abordaremos el problema de interpretación del lenguaje escrito usando gramáticas de estilos y formas lógicas. Con la finalidad de reducir la complejidad del procesamiento sintáctico/semántico e incorporar el conocimiento contextual en el proceso. Validaremos las estrategias con un prototipo de un módulo de asignación de descriptores para un sistema bibliográfico virtual.

Tabla de contenido:

1. Introducción: ¿Qué pretende el proyecto?

El proyecto pretende desarrollar una herramienta para interpretar documentos en español y extraer de descriptores. La implementación de dicha herramienta se confrontará con varios problemas históricos. Los problemas de procesar el lenguaje natural y de extraer información han sido atacados desde hace varias décadas [Moreno, A.], [Allen, J.], [Wilson, B.], como se reseña en el marco teórico de este trabajo. Sin embargo, todas las investigaciones realizadas no han sido suficientes para construir un sistema con un rendimiento cercano al humano al interpretar un lenguaje natural.

El lenguaje natural escapa a todos los esfuerzos de tratamiento computacional al parecer debido a que el conocimiento lingüístico está asociado, de formas sutiles y desconocidas, con el conocimiento contextual que tiene el hablante [Covington, M.] [Winograd, T.]. Este conocimiento resuelve en muchos casos la ambigüedad que no puede resolverse a nivel sintáctico y semántico. El problema con el conocimiento contextual es que se trata de cualquier conocimiento. Eso significa que tiene, primero que existir una forma de describir el mundo y segundo que esa forma de describir el mundo tiene que asociarse de alguna manera con el conocimiento lingüístico específico.

Integrar el conocimiento lingüístico y el conocimiento del mundo para tratar el lenguaje natural ha sido y será el problema general de cualquier sistema de procesamiento del lenguaje natural.

El desarrollo histórico del tratamiento de este problema destaca la iniciativa de usar el procesamiento del lenguaje natural desde los inicios de la computación. Con el advenimiento de los computadores, se ha tenido la idea de usarlos para obtener sistemas de extracción de información rápidos e inteligentes. Un ejemplo de esta situación son las bibliotecas, muchas de las cuales tienen ciertamente un problema de almacenaje y extracción de información. Las tareas como la catalogación y la administración en general, han sido controladas con éxito por las computadoras [van Rijsbergen. C.J.].

Desde los años 40 el problema de almacenaje y extracción de información ha atraído cada vez más atención, debido a que las cantidades de información han aumentado, y su acceso exacto y rápido se ha hecho más difícil. Como consecuencia de esto la información relevante generalmente no se consigue, y esto conduce a trabajo y esfuerzo duplicado.

Actualmente grandes esfuerzos de investigación están dirigidos a desarrollar y mejorar las tecnologías de recuperación de información, recuperación de texto y documentos, búsquedas inteligentes y extracción de conocimiento [Knight, K.] [Lewis y Sparck]. Todos estas tecnologías comparten el mismo problema: "el

procesamiento del lenguaje natural por parte del computador", sin duda se ha convertido en el objetivo hacia el cual se acerca el futuro de la computación.

Sin embargo, el procesamiento del lenguaje natural no solo es importante para atacar los problemas de extracción de información. También existe una necesidad cada vez mas primordial con respecto a las interfaces de los sistemas. Esto se debe a que los computadores participan en muchas tareas humanas, cada vez más individuos tienen acceso a la información a través de un computador y además la información también aumenta. Entonces se ha hecho necesario mejorar y simplificar la comunicación entre los sistemas y sus usuarios. Una alternativa es utilizar el lenguaje natural, lo cual representa una interfaz más sencilla y fácil de usar. Por tanto, el procesamiento del lenguaje natural es un componente importante de las interfaces de usuarios y los sistemas inteligentes.

Con estas expectativas la Inteligencia Artificial ha dirigido una parte de su trabajo hacia la programación de un computador para entender el lenguaje natural. Así se han realizados diversos procedimientos para procesar (y entender) el lenguaje natural. La lingüística teórica también ha aportado el fruto de su investigación y es entonces cuando en los sesenta se comienza a introducir la información lingüística en el procesamiento del lenguaje natural. Así se definió una área de conocimiento llamada Lingüística Computacional [Moreno, A.], apoyada por la Association for Computational Lingüistics (ACL).

Uno de los objetivos de la Lingüística Computacional es incorporar el conocimiento lingüístico a través de la simulación por el computador. La mayor parte del procesamiento del lenguaje en un computador es lineal, que no se corresponde con el procesamiento simultáneo y paralelo de nuestro cerebro. Esto ha estimulado la aparición de una aproximación inspirada en el funcionamiento neuronal, que probablemente representa una manera de simulación del cerebro por parte de la computadora [Moreno, A.].

Un enfoque predominante en los lingüistas computacionales es que el lenguaje es un proceso comunicativo donde emisor y receptor procesan determinada información en función de un conocimiento lingüístico y un conocimiento del mundo (pragmático) compartido [Winograd, T.]. Por tanto, para construir un sistema de procesamiento en una lengua natural, debe integrarse y utilizarse eficazmente diferentes tipos de conocimientos: sintáctico, semántico, discursivo y pragmático.

El estado actual de los sistemas de procesamiento del lenguaje natural muestran los logros alcanzados en estas décadas de investigación. Sin embargo, en comparación con el número de proyectos e inversiones realizadas, estos resultados no satisfacen las necesidades de estos sistemas. Debido a que no hay aplicaciones que se aproximen a la capacidad lingüística humana con una suficiente eficacia, robustez y fácil manejo.

Este trabajo inicialmente presenta un marco teórico, donde se resumen los diferentes modelos y tecnologías empleados para procesar documentos. Mostramos así la teoría en la que se ha basado el procesamiento del lenguaje natural desde su surgimiento. Luego en la sección de definición del problema, detallamos los intentos que se han realizado para interpretar un documento y obtener sus descriptores. También se presentan los problemas específicos y las hipótesis planteadas para el proyecto de tesis. Posteriormente se describe y justifica una estrategia metodológica para resolver el problema planteado. Finalizamos con las conclusiones derivadas de esta investigación.

2. Marco Teórico

En este marco teórico resumimos las diferentes tecnologías aplicadas en el procesamiento computacional de documentos escritos en lenguaje natural. Debido a que este es un problema que surgió en los años cuarenta, esta sección tiene un carácter histórico. Comenzamos explicando los primeros intentos por tratar el contenido de los documentos; hablamos entonces de la técnica de Recuperación de Información. Destacamos en especial la Minería de Texto, la Extracción de Información y Texto, debido a que son las tecnologías afines a la Recuperación de Información que están más relacionadas con nuestro problema. De los resultados de la aplicación de estas técnicas surge la necesidad de añadir el Procesamiento del Lenguaje Natural. Por lo tanto, a continuación mostramos la teoría en la que se ha basado el Procesamiento del Lenguaje Natural desde su surgimiento. Para esto, explicamos los tres modelos en los que se ha manifestado el procesamiento del lenguaje: Modelo Simbólico, Modelo Estadístico y Modelo Biológico. Culminamos con la comparación de estos tres modelos y con un resumen histórico que concluye con las tendencias actuales del Procesamiento del Lenguaje Natural.

2.1. Recuperación de Información.

La recuperación de información (IR Information Retrieval) es un término amplio, a menudo vagamente definido y en este contexto se refiere solamente a los sistemas automáticos de recuperación de información. Lancaster proporciona una definición: "Un sistema de recuperación de información no informa (es decir cambia el conocimiento) al usuario del propósito de su pregunta. Este informa simplemente de la existencia (o la no existencia) y paradero de documentos referentes a su petición" [Lancaster, F. W.].

Jacobs y Rau definen la recuperación de información como una tarea que, dado un conjunto de documentos y una consulta de usuarios, encuentra los documentos relevantes. Las aplicaciones de IR requieren velocidad, consistencia, precisión, y facilidad de uso en la recuperación de textos relevantes para satisfacer las consultas de los usuarios [Jacobs y Rau].

Según Moreno, la IR “se ocupa de tomar la consulta de un usuario a una base de datos y elegir entre todos los textos que se tienen archivados aquellos que mejor responda a las condiciones de búsqueda planteadas. Cuanto mayor sea el número de textos y más diversos sean los temas de los que tratan, más difícil será responder con exactitud”. De aquí surge la necesidad de “entender” realmente la pregunta y reconocer el contenido del documento.

Rijsbergen [van Rijsbergen. C.J.] resume las diferencias entre la recuperación de datos (DR Data Retrieval) y la recuperación de información (IR Information Retrieval), con algunas de las características que los distinguen en la Tabla 2.1.

Propiedad	**Data Retrieval (DR)**	**Information Retrieval (IR)**
Matching	*Match* exactos	*Match* parciales, el mejor *match*
Inferencia	Deducción	Inducción
Modelo	Determinístico	Probabilístico
Clasificación	*Monothetic*	*Polythetic*
Lenguaje de Consulta	Artificial	Natural
Especificación de Consulta	Completa	Incompleta
Itemes requeridos	*Matching*	Relevantes
Respuesta ante error	Sensible	Insensible

Tabla 2.1 Diferencias entre Recuperación de Datos y Recuperación de Información

Según la tabla anterior la recuperación de datos busca generalmente un *matching* exacto, es decir, se verifica si un *item* está o no está presente en el archivo. Este tipo de búsqueda puede, ser a veces de interés para la recuperación de información, pero generalmente se trata de hacer corresponder parcialmente estos *itemes* con la consulta y después se seleccionan los mejores.

La inferencia usada en la recuperación de datos es de la clase deductiva simple, es decir, aRb y bRc entonces aRc. En la recuperación de información es más común utilizar inferencia inductiva; las relaciones se especifican solamente con el grado de certeza o incertidumbre y por lo tanto la confianza en la inferencia es variable. Esta distinción conduce a describir a la recuperación de datos como una recuperación determinista y a la recuperación de información como probabilística.

Otra distinción que muestra la tabla 2.1 es en términos de la clasificación. En DR se esta mas interesado en una clasificación *monothetic*, donde las clases están definidas por los objetos que poseen atributos necesarios y suficientes para pertenecer a una clase. En IR tal clasificación en conjunto no es muy útil. En la clasificación *polythetic* cada individuo de una clase poseerá solamente una proporción de todos los atributos poseídos por todos los miembros de esa clase. Por lo tanto no hay atributos necesarios ni suficientes para ser miembro de una clase.

El lenguaje de consulta para DR generalmente es artificial, con una sintaxis y vocabulario restringido; en IR se prefiere utilizar el lenguaje natural. En DR la consulta es una especificación completa de lo que se desea, mientras que en IR generalmente es incompleta. Esta última diferencia se presenta debido a que en la IR se están buscando lo documentos relevantes, en lugar de *items* que correspondan exactamente. Por último la DR es más sensible al error en el sentido que, un error en el *matching* no recuperara el *item* deseado lo que implica una falla total del sistema. Los pequeños errores en los *matching* de IR generalmente no afectan de manera perceptible el funcionamiento del sistema.

Aunque la recuperación de información se puede subdividir de muchas maneras, hay tres campos de investigación principales, ellos son: análisis del contenido, estructuras de la información, y evaluación [van Rijsbergen. C.J.]. El primero tiene que describir abreviadamente el contenido de los documentos en una forma conveniente para el tratamiento a través del computador; el segundo se refiere a explotar relaciones entre los documentos para mejorar la eficacia de las estrategias de la extracción; el tercero con la medida de la eficacia de la extracción. Para tener un resumen del desarrollo histórico de IR se pueden consultar los textos de [Cleverdon, C.W.] y [Salton, G.].

Un sistema de IR típico se presenta en la figura 2.1. El diagrama muestra tres componentes: entrada de información, procesamiento y salida.

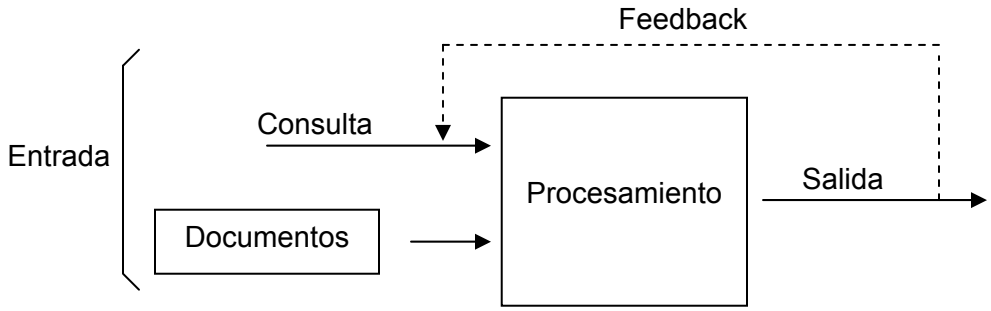

Figura 2.1. Un sistema de Recuperación de Información.

Comenzando con la "Entrada" de información. El problema principal aquí es obtener una representación de cada documento y preguntar usando un computador. Generalmente la mayoría de los sistemas de extracción computarizados almacenan solamente una representación del documento (o de la pregunta). Un representante del documento podía, por ejemplo, ser una lista de las palabras extraídas consideradas significativas. Mucho mejor es hacer que el computador procese el lenguaje natural, un acercamiento alternativo es tener un lenguaje artificial dentro del cual todas las interrogaciones y documentos puedan ser formulados.

El componente denominado "Procesamiento", esta relacionado con el proceso de recuperación. Este puede requerir la estructuración de la información de alguna manera apropiada, tal como la clasificación. En [van Rijsbergen. C.J.] se presenta una introducción a los métodos de recuperación de información. Uno de estos métodos son la clasificación [Sparck, J.] [Cormack, R.M.], es necesario conocer la descripción de la clase de datos para los cuales los métodos de clasificación son apropiados. Los datos consisten de objetos y sus descripciones correspondientes. Los objetos pueden ser documentos, palabras claves, caracteres escritos a mano, etc.

Finalmente, se muestra el componente de la salida. Debido a que la mayoría de los sistemas de IR tenían como objetivo simplemente una búsqueda bibliográfica, el resultado del procesamiento de la consulta es una salida simple. Esta salida consiste del número de documentos recuperados (documentos relevantes para la consulta) y sus correspondientes citas.

Luego de entender la idea general de los sistemas de IR, parece importante destacar lo siguiente. Según van Rijsbergen nunca no se había asumido que un sistema de recuperación debería procurar "entender" el contenido de un documento. Los documentos se juzgaban como relevantes en base a una descripción superficial. En el libro de van Rijsbergen se da una introducción sobre este tratamiento, y se pone de manifiesto la necesidad de un procesamiento del lenguaje natural [van Rijsbergen. C.J.].

De esta manera, a partir de la década de los noventa se ha presentado un crecimiento de los experimentos en el uso de técnicas de lenguaje natural para incrementar los resultados y la precisión en la recuperación de información. Algunos ejemplos de estos experimentos se encuentran en [Croft y Lewis] [Fagan, J.]. Jacobs y Rau creen que es solamente cuestión de tiempo, para que estos esfuerzos se incorporen a la próxima generación de sistemas de recuperación de información [Jacobs y Rau]. Incluso algunos investigadores como Knight consideran a los sistemas de IR actuales como aplicaciones de lenguaje natural que están al alcance de una gran cantidad de usuarios [Knight, K.].

A partir de las técnicas básicas de recuperación de información, se han originado diferentes tecnologías entre ellas las denominadas Minería de Texto y Extracción de Información.

2.1.1. Minería de Texto

Se trata de un grupo de herramientas destinadas a la recuperación de información en los sistemas. Este grupo de herramientas implementa una tecnología llamada "Minería" sirve para identificar y extraer valores importantes de los datos, descubrir conocimiento, informaciones ocultas, asociaciones, patrones y características [IBM-1]. Existen dos enfoques, la minería de datos y la minería de texto. La minería inteligente de datos permite minar la estructura de los datos almacenados en una base de datos convencional. En cambio, en la minería inteligente de texto la fuente de información proviene de textos en lenguaje natural, tales como correspondencia de los clientes, servicios de noticias, correo electrónico, páginas web, etc.

La minería de texto permite extraer patrones de texto, documentos organizados por sujetos, temas predominantes en la colección de documentos y búsqueda de documentos relevantes [IBM-2]. Los servicios que presta la minería de texto son diversas:

- Asignar documentos a categorías predefinidas, obteniendo una lista de nombres de categorías y niveles.
- Dividir documentos en grupos, llamados "clusters" para facilitar el proceso de exploración y encontrar información similar o relacionada. Aquí se pueden identificar relaciones ocultas entre los documentos, y se descubren documentos repetidos.
- Extraer características (nombres, terminologías, abreviaciones, etc.) automáticamente.
- Identificar el idioma del texto, indexar por el lenguaje y restringir las búsquedas en un lenguaje particular.
- Búsquedas por texto, realizan un análisis lingüístico para procesar las consultas en lenguaje natural y procesar los diccionarios.

La minería de texto, es considerada una técnica de descubrimiento del conocimiento, las cuales automáticamente encuentran propiedades fundamentales e información útil [Toshinori, M]. En particular la minería es muy usada en las aplicaciones de computación actuales, por ejemplo procesadores de palabras, herramientas de comunicación en Internet, búsquedas en el web, y herramientas de traducción.

Por otra parte, hay una tendencia a usar la minería con otros propósitos. Knight dice que la construcción de aplicaciones de minería en procesamiento de lenguaje natural posee desafíos extraordinarios, pero también ofrece grandes recompensas. En particular menciona varios trabajos en los cuales se ha aplicado minería de texto para obtener y descubrir información lingüística, uso del lenguaje y semántica de palabras en varios contextos. Estos resultados han sido importantes porque representa un conocimiento que pueden alimentar a los sistemas de procesamiento de lenguaje natural. Con esto se puede mejorar la solución de los problemas de ambigüedad y automatizar las tareas manuales de estos sistemas, tales como la generación de reglas gramaticales y la construcción de diccionarios [Knight, K.].

2.1.2. Extracción de Información

La extracción de información (IE Information Extraction) trata colecciones de textos. Luego los transforma en información que es digerida y analizada mas fácilmente. Esto identifica los fragmentos de textos relevantes, extrae la información relevante de los fragmentos, y con estas piezas organiza la información requerida en una estructura coherente. Se trata de reconocer la información importante contenida en los documentos y trasladarla a un formato predefinido para que pueda ser tratada y recuperada con mayor facilidad.

El objetivo de los investigadores de IE es construir sistemas que encuentren items que puedan ser de interés para el análisis humano a partir de documentos. Además de la información relevante deben conseguirse las relaciones entre ellos,

mientras que se ignora la información irrelevante y extraña. Hoy día, sin embargo, los sistemas de IE tratan solamente con tipos específicos de textos y solo tienen buenos resultados en algunos componentes [Cowie y Lehnert].

Según Elmasri y Navathe hay problemas no resueltos en la extracción de texto. Aunque por largo tiempo la extracción de texto se ha utilizado en las aplicaciones de negocios y en los sistemas bibliotecarios, los diseñadores de sistemas de extracción de texto se siguen enfrentando a estos problemas de: indicación de frases, empleo de diccionarios de sinónimos (localizar un tesauro adecuado para el dominio de interés) y resolución de la ambigüedad [Fagan, J] [Krovetz y Croft] [Salton y Buckley].

Por otra parte, Elmasri y Navathe también determinan tres enfoques básicos para interpretar el lenguaje natural en la extracción de texto [Elmasri y Navathe]:

(1) Extracción de palabras claves y comparación de patrones: En los sistemas de palabras claves, el programa relaciona palabras del lenguaje natural con campos específicos de una base de datos, y el creador de la aplicación define los vínculos, Sin embargo, la comparación de patrones sin una base gramatical no tiene sino una utilidad limitada.
(2) Análisis sintáctico: El análisis sintáctico convierte una frase expresada en un lenguaje natural. Potencialmente ambigua, a un formato interno que deberá representar con precisión la consulta que desea hacer el usuario. Hay dos posibles variaciones de este enfoque: una basada en una gramática del lenguaje natural y la otra basada en la semántica del lenguaje en términos de un léxico y una serie de reglas de producción.
(3) Transformaciones de consultas: Se utiliza una base de conocimientos o un "modelo del mundo" con representaciones canónicas de enunciados de un cierto dominio de aplicación, junto con conocimientos lingüísticos y conocimientos de transformación, para transformar consultas expresadas en lenguaje natural en un nivel conceptual a consultas de base de datos expresadas en un lenguaje específico.

Se puede realizar la combinación de estos tres enfoques básicos, con el fin de compensar las limitaciones de cada uno de ellos. Una aplicación puede realizar una extracción de texto mas adecuada si, realiza un análisis sintáctico, aplica transformaciones en base a un modelo del dominio, y además esta apoyada de patrones del lenguaje. La tendencia actual de los sistemas con tecnologías de la información que procesan el lenguaje natural, es precisamente los modelos híbridos.

2.2. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (NLP Natural Language Processing), originalmente desarrollado a comienzos de la Guerra Fría [Locke y Booth] como el

mecanismo que usaban los físicos Soviéticos para la traducción de documentos, es uno de los primeros objetivos computacionales mas investigados. Estos esfuerzos prematuros, por analizar y modelar el lenguaje humano, fueron caracterizados por una técnica sin conocimiento lingüístico y por el bajo rendimiento computacional de la época.

Según Covington "El Procesamiento de lenguaje natural (NLP) es el uso de computadoras para entender lenguajes (naturales) humanos tales como ingles, francés o japonés. Por entender no se quiere decir que el computador tenga pensamientos, sentimientos y conocimientos humanizados, sino que el computador pueda reconocer y usar información expresada en lenguaje humano" [Covington, M.].

Manaris y Slator definen a un sistema de NLP como aquel que encapsula un modelo del lenguaje natural en algoritmos apropiados y eficientes. En donde las técnicas de modelado están ampliamente relacionadas con eventos en muchos otros campos, incluyendo [Manaris y Slator]:

- Ciencia de la computación, la cual provee métodos para representar modelos, diseñar e implementar algoritmos para herramientas de software.
- Lingüística, la cual contribuye con nuevos modelos lingüísticos y procesos.
- Matemática, la cual identifica modelos formales y métodos.
- Neurociencia, la cual explora los mecanismos mentales y otro tipo de actividades físicas.

Entre estos campos, la lingüística ha aportado el conocimiento lingüístico de las lenguas naturales. Este conocimiento dentro de un sistema de NLP puede ser dividido en niveles definidos en términos de la característica declarativa (qué) y procedural (cómo), tal como se muestra en la Tabla 2.2. [Manaris y Slator].
Como se puede observar en la tabla 2.2., el conocimiento lingüístico se puede organizar en diferentes niveles o componentes, ya que la estructura de cualquier lenguaje humano se puede dividir naturalmente en estos niveles [Covington, M.]:

- **Fonológico**: la fonología estudia como los sonidos (sonidos hablados) son usados en el lenguaje. Cada lenguaje tiene un alfabeto de sonidos que se distinguen, ellos son llamados fonemas. Este nivel trata de las realizaciones acústicas, por tanto solo aparecen en los sistemas de reconocimiento del habla. Tecnológicamente, el tratamiento del habla por parte del computador esta un poco separado del resto del procesamiento del lenguaje natural. Debido a que este tratamiento del habla tiene relación con el análisis de la forma de la onda del sonido y el reconocimiento de patrones. Mientras que el resto de los niveles dependen de una programación simbólica y un razonamiento automatizado.
- **Morfológico**: la morfología es la rama de la lingüística que se preocupa por la descripción de la estructura de las palabras y los procesos de formación de las palabras. La idea general es que los morfemas individuales, pueden ser combinadas para formar palabras. Hay tres procesos diferentes en la formación

de palabras:

- La inflexión: La Morfología Inflexional se preocupa por las relaciones gramaticales tales como el plural, tiempo pasado, y la posesión. Los afijos de inflexión no cambian la categoría sintáctica de las raíces a las que ellos están conectados. Así por ejemplo, tanto árbol como árboles (la raíz "árbol" mas el afijo plural "es") son nombres.
- La derivación: La Morfología Derivacional describe como son creadas nuevas palabras con la ayuda de afijos. Por ejemplo, el adjetivo nacional se deriva del sustantivo nación.
- Composición: La Composición se preocupa por la construcción de palabras nuevas combinando morfemas libres, como en paraguas, de "para" y "agua".

La morfología es un componente primordial para aquellas lenguas ricas en formas flexionadas (como el español o el alemán). La morfología es útil para evitar la expansión innecesaria de formas completamente flexionadas en el diccionario, cada palabra tiene que ser codificada con sus rasgos morfosintácticos en el lexicón.

	Características	
Nivel	Declarativo (qué)	Procedural (cómo)
Fonológico	Sonidos hablados	Formar morfemas[1]
Morfológico (Léxico)	Unidades de palabras, Palabras	Formar palabras, Derivar unidades de significado.
Sintáctico	Roles estructurales de palabras (o colección de palabras)	Formar oraciones
Semántico	Significado independiente del contexto	Derivar significado de oraciones
Discurso	Roles estructurales de oraciones (o colección de oraciones)	Formar diálogos
Pragmático	Significado dependiente del contexto	Derivar significado de oraciones relativo al discurso circundante

Tabla 2.2. Niveles de conocimiento en el procesamiento del Lenguaje Natural

- **Sintáctico**: la sintaxis, o construcción de oraciones, es el nivel mas bajo en el cual el lenguaje humano es constantemente creativo. Noam Chomsky (1957) fue el primero en hablar sobre este punto. El introdujo las "gramáticas generativas", cuyas oraciones son descritas por reglas dadas, en lugar de listar las oraciones y sus estructuras directamente, se construyen las oraciones a partir de estas reglas.

  El conocimiento sintáctico es un componente básico de cualquier sistema de

[1] Morfemas: son las unidades distintivas mínimas de la gramática. Hay dos clases de morfemas: *Formas Libres* la cual puede ocurrir como palabras separadas, y *Formas de Salto*, que no pueden ocurrir como palabras en si mismas. Los últimos son generalmente llamados Afijos. Por ejemplo, la palabra en inglés "unselfish" se compone de tres morfemas, "un", "self", y "ish". El "self" es una forma libre mientras "un" y "ish" son formas de salto. En el particular, "un" es aquí un prefijo, "ish" es un sufijo y "self" es una raíz.

NLP, pues se encarga de reconocer las oraciones gramaticales y asignarles una estructura. El reconocimiento de la estructura de las oraciones por parte del computador es llevado a cabo por un algoritmo llamado “parsing”.

- **Semántico**: la semántica, o significado, es el nivel en el cual el lenguaje hace contacto con el mundo real. Se trata de la primera tarea del componente interpretativo, la cual consiste en asignar un significado a cada una de las oraciones analizadas independientemente del contexto. Se realiza lo que se denomina composición semántica, que es la composición de significados de palabras para formar significados de oraciones. Por ejemplo “Juan ama a María”, se forma del significado de “Juan”, “ama” y “María”, y se puede representar como formulas lógicas así: ama(Juan,María). La semántica oracional es una parte imprescindible de cualquier sistema, ya que sin ella no podríamos asignar significado a las estructuras analizadas.
- **Discurso**: aquí se tratan los aspectos de interpretación afectados por las oraciones emitidas anteriormente. En este nivel se almacena el conocimiento que permite relacionar entre si el significado de las oraciones aisladas e integrarlo para formar unidades mayores. En concreto, este conocimiento se utiliza para interpretar los pronombres anafóricos[2], resolver los elementos elididos[3] y los aspectos temporales. Este componente es necesario para que los sistemas tengan conocimiento del contexto comunicativo en el que se están produciendo los mensajes y tiene en cuenta aspectos pragmáticos como las intenciones del emisor y del receptor [Moreno, A.].
- **Pragmático**: se refiere al uso del lenguaje en el contexto. En general la pragmática incluye aspectos del conocimiento conceptual del mundo que van mas allá de las condiciones reales literales de cada oración. Este conocimiento lo tienen en cuenta los hablantes cuando se comunican mediante una lengua. Les sirve para comprender mucha información sobreentendida pero no expresada explícitamente en las oraciones. Mientras la sintaxis y semántica estudia las oraciones, la pragmática estudia “las acciones del discurso” y las situaciones en las cuales el lenguaje es usado.

  Muchas palabras y oraciones pueden ser ambiguas y tener mas de un significado, su significado puede ser falso o producir implicaciones falsas. El significado depende de los principios que usan las personas cuando hablan (por ejemplo ser relevantes y hacer énfasis en las oraciones verdaderas [Grice, H.]). En este sentido la pragmática tiene dos conceptos importantes: la implicación y la presuposición de las oraciones. La implicación de una oración comprende la información que no es parte de su significado, pero que debe ser inferida por un oyente razonable. Las presuposiciones de una oración son las cosas que deben ser verdaderas para que la oración sea verdadera o falsa. Es decir, en base a

[2] Anáfora: es una expresión que se refiere a una previa expresión de un discurso en lenguaje natural. Generalmente usa un pronombre para referirse a personas, lugares o cosas previamente mencionadas. Por ejemplo “María murió. Ella estaba muy vieja”, ella se refiere a María.

[3] Elipsis: se refiere a las situaciones cuyas oraciones son abreviadas o eliminan un constituyente, dejando parte de ellas para ser entendidas por el contexto. Por ejemplo cuando se pregunta “¿Cuál es tu nombre?”, y se contesta “Juan Pérez” esta es una forma elíptica de “Mi nombre es Juan Pérez”.

las presuposiciones (conocimiento verdadero de un dominio) las personas interpretan las oraciones y derivan conocimiento (implicaciones) que pueden ser o no verdaderos.

El conocimiento lingüístico se incorporo a los sistema de NLP a partir de los años sesenta, y se convierte en uno de sus componentes importantes. A parir de ese momento se definió una área de conocimiento llamada Lingüística Computacional (CL Computational Linguistics) apoyada por la Asociación para la Lingüística Computacional (ACL Association for Computational Linguistics).

2.2.1. Lingüística Computacional

Según Moreno la Lingüística Computacional es una disciplina que trata básicamente de dos cosas: lenguas naturales y computadoras. Muchas líneas de investigación comparten ambos objetivos aunque desde perspectivas diferentes. Como siempre hay que enfrentarse con el objeto de estudio y con la delimitación de las terminologías de las ciencias, hay que dejar claro que la lingüística computacional es equivalente al NLP, y no es igual a la lingüística informática[4] y a la ingeniería lingüística[5] [Moreno, A.].

Siguiendo a Grishman, se puede definir la lingüística computacional como "el estudio de los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales" [Grishman, R.]. Una definición equivalente proporciona Allen para el Procesamiento de lenguaje natural "El objetivo de esta investigación es crear modelos computacionales del lenguaje lo suficientemente detallados que permitan escribir programas informáticos que realicen las diferentes tareas en donde interviene el lenguaje natural" [Allen, J.]. Por tanto, la CL y el NLP tratan de lo mismo: del desarrollo de programas de ordenador que simulan la capacidad lingüística humana.

La Inteligencia Artificial (AI Artificial Intelligence) se encarga de codificar en un programa facultades cognitivas como la inferencia, la toma de decisiones, la adquisición de conocimiento, etc. En este sentido, la CL es una parte integrante de la AI, de la misma forma que para muchos lingüistas la Lingüística es parte de la Sicología por tratar una de las capacidades cognitivas por excelencia, el lenguaje.

Según Moreno, la Lingüística Computacional trata de la construcción de sistemas informáticos que procesen realmente estructuras lingüísticas y cuyo objetivo sea la simulación de la capacidad lingüística humana, independientemente de su carácter

---

[4] Lingüística informática: es una disciplina que abarca el uso de computadores con relación al lenguaje y a las lenguas. Incluye todo tipo de herramientas que ayuden al estudio de las lenguas y de la lingüística. La lingüística Computacional es una parte de la lingüística informática [Moreno, A.].

[5] Ingeniería lingüística: se refieren a las aplicaciones potencialmente comerciales que implican el uso de nuevas tecnologías. Incluye la edición electrónica (diccionarios, libros), los productos multimedia, etc. [Moreno, A.].

comercial o de investigación básica [Moreno, A.].

Este trabajo de investigación usara los términos NLP y CL indistintamente. El termino de NLP, aparecerá con mas frecuencia pues suele ser mejor entendido.

Para terminar de perfilar el contenido de la disciplina de Lingüística Computacional, es importante conocer las principales aplicaciones prácticas. Moreno presenta una clasificación de estos tipos de aplicaciones [Moreno, A.]:

1. Sistemas que tratan de emular la capacidad humana de procesar lenguas naturales. Dentro de este grupo las mas importantes son: Traducción automática, Recuperación y extracción de Información, Interfaces hombre-maquina.
2. Sistemas que ayudan en las tareas lingüísticas. Este grupo esta formado por herramientas que pueden ser utilizadas por los lingüistas para facilitarles ciertas tareas complejas. Algunas aplicaciones de este tipo son: Herramientas de análisis textual, Herramientas de manejo de corpus[6], Bases de datos lexicográficas.
3. Programas de ayuda a la escritura y composición textual. Las aplicaciones comprendidas en este grupo han sido ampliamente desarrolladas y cualquier usuario habitual de un procesador de texto esta familiarizado con ellas: Correctores ortográficos, Correctores sintácticos y de estilo.
4. Enseñanza asistida por computador. Este es un campo de aplicación en continua expansión y que tiene varias vertientes. La mas importante es la de los programas educativos para la enseñanza de las lenguas extranjeras.

2.2.2. El Lenguaje desde el punto de vista Científico.

La definición del lenguaje desde el punto de vista científico ha llevado a muchos lingüistas a estar de acuerdo en diferentes puntos. Covington presenta los mas importantes [Covington, M.]:

- *El lenguaje es forma y no sustancia*. Esto quiere decir, que el lenguaje no es un conjunto de pronunciaciones o comportamientos- es un sistema de reglas que determina el comportamiento. Otra manera de decir esto es distinguir entre la competencia del hablante (el sistema) y su rendimiento (su comportamiento observable). Esta distinción reconoce que las pronunciaciones accidentales, oraciones interrumpidas, etc. no son realmente instancias del lenguaje que habla la persona, sino son derivaciones del lenguaje.
- *El lenguaje es arbitrario*. Un lenguaje es un conjunto de símbolos que las personas acuerdan usar de una manera especifica.

[6] Corpus: es una colección de datos lingüísticos, normalmente formados por varios textos. Esta gran cantidad de textos en lenguaje natural son usados para acumular estadísticas de textos en lenguaje natural.

- *Todos los lenguajes humanos usan dualidad de patrones*. En donde las palabras son cadenas de sonidos, y las oraciones son cadenas de palabras. Las palabras tienen significado, los sonidos en si mismos no.
- *Todos los lenguajes son casi igualmente complicados*, excepto por el tamaño del vocabulario. Los lenguajes cambian constantemente, pero cada cambio es lento. Un lenguaje evoluciona en una dirección particular en cientos de años.
- *Todo el mundo habla su propio lenguaje*. El idioma español que habla una persona no es completamente igual al idioma español que habla su padre. Esto se debe a que la manera en que el lenguaje es aprendido. Según el aprendizaje del lenguaje se producen pequeñas diferencias entre individuos, y grandes e inevitables diferencias entre los grupos sociales.

2.2.3. Problemas en el uso del Lenguaje Natural.

El conocimiento del mundo es un factor importante en los sistemas de NLP. Por tanto, un sistema NLP debe colocar limites a la necesidad de conocimiento externo y de la experiencia humana. Covington dice que además de lo anterior, el NLP depende de otras dos factores: La primera se refiere a que al poder de las computadoras. La aparición de microcomputadores en 1980 ha marcado la diferencia. Previamente, el NLP fue tan costoso que las personas aceptarían cualquier resultado perfecto, lo cual nunca fue alcanzado. Esta situación ha cambiado y las aplicaciones aunque imperfectas son más económicas, y los usuarios encuentran buenos usos para ellas.

El segundo factor y quizás la más importante, es que el NLP depende del conocimiento exacto de como el lenguaje humano trabaja -lo cual, ahora, no se conoce suficientemente. Hasta hace pocos años, el lenguaje fue estudiado casi exclusivamente con el propósito de enseñarle un idioma a otros humanos. El principio que esta detrás de todos los lenguajes humanos fue ignorado. Por otra parte, la ciencia de la lingüística tiene solamente unas pocas décadas de antigüedad, y no hay todavía consensos en relación a algunos hechos básicos.

Los sistemas de NLP deben atacar una variedad de problemas relacionados con el lenguaje natural [Manaris y Slator]:

- *Inexactitud*, incluyendo errores ortográficos, signos de puntuación incorrectos, palabras transpuestas, y oraciones agramaticales.
- *Incompletitud*, incluyendo construcciones elípticas, anáforas, etc.
- *Imprecisión*, incluyendo el uso de términos relativos sin un punto especifico de referencia y el uso de términos cualitativos
- *Ambigüedad*, debido a que pueden surgir múltiples interpretaciones en cualquier nivel del conocimiento lingüístico (ver tabla 2.2). la ambigüedad puede ser resuelta usando el conocimiento de un nivel mas alto.

2.3. Modelos del Procesamiento del Lenguaje Natural

Los modelos y métodos de NLP pueden ser clasificados en: simbólicos, empíricos o estadísticos, conexionistas y los enfoques híbridos. Los dos primeros son llamados modelos matemáticos del lenguaje. El enfoque simbólico esta basado en el conocimiento, emplea reglas y algoritmos que operan con estructuras de datos simbólicos que representan el conocimiento del lenguaje natural. El enfoque empírica o estadística involucra colecciones de muestras del lenguaje (corpus), las cuales son etiquetadas y usadas para crear modelos estadísticos para NLP. La técnica conexionista usa redes neuronales para representar el conocimiento lingüístico. Por otra parte, las técnicas híbridas combinan uno o mas de los modelos anteriores, con el fin de complementar las ventajas de cada uno y resolver problemas de dominios y aplicaciones específicos. Los primeros tres enfoques son explicados a continuación.

2.3.1. Modelos simbólicos

Los sistemas simbólicos se basan en la manipulación de símbolos, ellos fueron concebidos por los matemáticos para captar de manera rigurosa y sistemática la demostración de teoremas matemáticos y lógicos. Dentro de la lingüística, Chomsky fue el primero en introducir de manera sistemática el paradigma lógico formal.

Típicamente las reglas de inferencia en un sistema formal permiten concentrarse en la sintaxis del modelo, independientemente de su interpretación. Muchos lingüistas piensan que el lenguaje tiene una naturaleza regular o lógica, y eso es lo que tratan de reflejar en sus gramáticas formales. En general, estas gramáticas han demostrado ser eficaces en la descripción y explicación de fenómenos relacionados con la competencia[7] [Moreno, A.].

2.3.1.1. Fundamentos teóricos: gramáticas formales

Una gramática formal es una especificación rigurosa y explícita de la estructura de una lengua. Esta es escrita con un formalismo gramatical, es decir, una lengua artificial creada para describir lenguas naturales. Su uso se debe a que es un lenguaje bien definido, riguroso, facilita la evaluación de hipótesis, y permite desarrollar predicciones.

Existen diferentes tipos de gramáticas que están formalizadas rigurosamente. Entre ellas tenemos: las gramáticas generativas, gramáticas categoriales, gramáticas de dependencia, gramáticas de cadenas lingüísticas de Harris y gramáticas de adjunción de árboles [Winograd, T.] [Grishman, R.] [Moreno, A.]. Sin embargo las mas conocidas son las gramáticas generativas, también conocidas

[7] Competencia: se refiere al conocimiento que cada hablante tiene de su lengua materna

como gramáticas de estructura de frase o sintagmáticas, propuestas por Chomsky [Chomsky, N.].

Las gramáticas generativas están constituidas por un conjunto de reglas generativas (o derivadas) que asignan explícitamente la estructura interna de las oraciones. Dichas reglas, llamadas reglas de reescritura, operan sobre los conjuntos de elementos no terminales y terminales. Tanto las gramáticas transformacionales como las de unificación son gramáticas generativas. Según Bach en 1974, cualquier gramática que defina precisa y explícitamente las oraciones de una lengua es una gramática generativa [Bach, E.]. Son las más extendidas en Lingüística Computacional.

Chomsky estableció una clasificación de tipos de gramáticas, aplicadas a las gramáticas generativas o sintagmáticas, que se ha hecho famosa con el nombre de su creador, la Jerarquía de Chomsky. Esta jerarquía esta organizada de acuerdo con el "poder generativo débil". El concepto de poder generativo o formal se utiliza para referirse a la capacidad de predicción de una gramática. En concreto, el poder generativo débil concierne a que tipo de oraciones la gramática puede reconocer como gramaticales.

Hay cuatro tipos de gramáticas generativas, cada uno definido por la clase de reglas que contiene. La tabla 2.3 muestra la jerarquía de Chomsky para clasificar las gramáticas generativas.

Tipo	Gramáticas	Restricciones a la forma de las reglas	Lenguas	Autómatas
0	Irrestrictas	Ninguna: $\alpha 1 ... \alpha N \rightarrow \beta 1 ... \beta N$	Enumerables recursivamente	Maquinas de Turing
1	Dependientes del contexto	La parte derecha contiene como mínimo los símbolos de la parte izquierda: $\alpha \rightarrow \beta$ / X _ Y o alternativamente: x α z $\rightarrow$ x β z...	Dependientes del contexto	Autómatas linealmente finitos
2	Independientes de Contexto	La parte izquierda solo puede tener un símbolo: $\alpha \rightarrow \beta$...	Independientes del contexto	Autómatas PDS (Push Down Store)
3	Regulares o de estados finitos	La regla solo puede tener estas dos formas: A $\rightarrow$ tB A $\rightarrow$ t	Regulares	Autómatas finitos

Tabla 2.3. Jerarquía de Chomsky [Moreno, A.]

Esta clasificación es completamente teórica, pues no existen tipos puros de gramáticas. En la practica, las gramáticas formales se van modificando según las

necesidades particulares [Moreno, A.]. Como consecuencia de esto, no se puede decidir fácilmente a que tipo pertenece una gramática. He aquí la distancia entre las demostraciones teóricas y las realizaciones practicas, la cual se asemeja a las diferencias entre los lingüistas teóricos[8] y los lingüistas computacionales.

En lingüística computacional se han obtenido ciertos resultados y aplicaciones con las siguientes gramáticas:

A. Gramáticas regulares o de estados finitos

Las gramáticas regulares o también llamadas red de transición, están formadas por nodos o estados (representados por círculos) y por arcos (representados por flechas) etiquetados. Cada arco representa una transición entre dos estados. Hay dos clases especiales de estados: los estados iniciales (marcados con una pequeña flecha) que son los únicos que no reciben flechas procedentes de otros arcos, y los estados finales (representados con doble circulo) que son los únicos de los que no parten transiciones a otros estados. No es posible tratar la recursividad con una gramática regular, ya que la única información que maneja es el estado en que se encuentra.

Esta gramática de estados finitos se ha aplicado a la morfología y al reconocimiento léxico. Debido a que las reglas de flexión forman un conjunto casi cerrado y mucho más pequeño que las reglas de sintaxis en cualquier lengua. En definitiva se ha usado en muchas tareas "finitas" del lenguaje, proporcionando un método muy eficiente para el computador [Roche y Schabes].

B. Gramáticas Independientes del Contexto.

El término Gramática Independientes del Contexto (CFG Context Free Grammar) es un modelo particular para describir la sintaxis del lenguaje. Los CFGs pueden ser usados para describir cualquier lenguaje (natural o artificial) conforme a algunas limitaciones bastante básicas. Las limitaciones se refieren a cómo las cláusulas se generan, y dicen rígidamente que las cláusulas dependientes deben estar adyacentes al componente del cual dependen. La mayoría de los lenguajes naturales parecen seguir este comportamiento, con la excepción posible del idioma alemán suizo [Gazdar y Mellish]. La mayoría de las teorías contemporáneas de sintaxis de lenguaje natural se derivan del esquema de CFG.

Como se muestra en la jerarquía de Chomsky (ver tabla 2.3.), la CFG es una gramática de tipo2, que están formadas por un conjunto de reglas y un conjunto de entradas léxicas (o lexicón). Mediante estas reglas se pueden describir la

[8] La lingüística teórica: se centra en analizar la competencia de los hablantes, utiliza principalmente la introspección para obtener sus datos y suelen llegar a sus conclusiones mediante métodos deductivos. Sus principales objetivos son conseguir una teoría gramatical, simple, restringida y que de cuenta de los universales lingüísticos.

estructura sintáctica de múltiples fenómenos de las lenguas naturales. Estas gramáticas son capaces de reconocer y generar oraciones.

Un CFG es una descripción formal de la sintaxis de un lenguaje. Específicamente, la descripción se da como un conjunto de "Reglas de Producción", que definen las oraciones bien formadas del lenguaje. Las reglas, por supuesto, son en sí mismas también escritas en un lenguaje formal. Como todos los idiomas formales, este es uno definido por un vocabulario y una sintaxis. En la tabla 2.4. se muestra la definición formal de las CFG.

Gramática Independiente del Contexto	
El vocabulario	Las reglas contienen tres tipos de símbolos: • **No Terminales**: Corresponde a los componentes del lenguaje a describir. Uno de los no terminales tiene una posición especial. A este se le llama el símbolo distintivo. • **Terminales**: Corresponde a las palabras del lenguaje a describir. • → : (El símbolo de la flecha) Delimita el lado izquierdo de una regla de su lado derecho.
La sintaxis	Las oraciones en el lenguaje del CFG son las reglas de producción (la sintaxis aquí se refiere al formato de las reglas en sí mismas, no al lenguaje que ellos describen. Una regla de producción tiene las siguientes propiedades: 1. Se compone de un lado izquierdo (LHS Left. hand side) y un lado derecho (RHS Right hand side) separados por una flecha: LHS → RHS 2. El LHS se compone de un solo símbolo no terminal. 3. 3. El RHS consiste o de uno o más no terminales, o solo terminal.

Tabla 2.4. Gramáticas Independientes del Contexto

Estas gramáticas proporcionan la estructura jerárquica interna de las oraciones. Se pueden describir construcciones recursivas que no podían ser tratadas con las gramáticas regulares. También permite expresar la alternancia y la opcionalidad. Además las gramáticas independientes del contexto tienen propiedades formales que facilitan el diseño de algoritmos de parsing. Sin embargo hay problemas con el tratamiento de ciertos fenómenos lingüísticos como los constituyentes discontinuos[9], la subcategorización[10] y la concordancia[11]. Moreno afirma que en la

[9] Constituyentes discontinuos son constituyentes que se pueden encontrar en mas de una posición estructural [Moreno, A.].

[10] Subcategorización: es un fenómeno básicamente léxico-semántico en donde la estructura oracional se predice en función de la semántica verbal. Tiene una importancia esencial en la sintaxis porque especifica las posibilidades de combinación de las palabras. Los verbos y algunos adjetivo admiten una estructura de complementos, la subcategorización se refiere al numero y a la categoría de los complementos de cada verbo.

[11] Concordancia: es un fenómeno de muchos lenguajes en donde las palabras toman ciertas inflexiones dependiendo de relación que guardan con las otras palabras de una oración. Esta relación se refiere al genero, numero. Un ejemplo simple ocurre con los verbos en tercera persona del singular y sus sujetos: "ella baila", "tu bailas".

practica, ningún sistema de PLN de cierta cobertura utiliza la versión pura de este tipo de gramáticas [Moreno, A.].

C. Gramáticas de Unificación y Rasgos.

Las restricciones más comunes que aparecen en las gramáticas independientes del contexto son: los fenómenos de concordancia y subcategorización. Las gramáticas de unificación y rasgos logra tratar ambos casos [Kay, M.], por eso son consideradas como el modelo computacional más completo y restringido al mismo tiempo conocido hasta la fecha.

Estas gramáticas se caracterizan por hacer complejas descripciones formales mediante el uso de rasgos y por utilizar una operación general para la combinación y comprobación de la información gramatical, conocida como unificación [Shieber, S.].

La estructura de rasgos es el mecanismo básico de representación de la información de las unidades lingüísticas. Por tanto cada elemento de información (unidad lingüística) tiene asociado una estructura de rasgos. En donde un rasgo es un par compuesto por un atributo y un valor. El atributo lleva el nombre que identifica el rasgo. Por ejemplo "número = plural" es un ejemplo de rasgo (atributo número, valor plural). Los valores de los rasgos pueden ser valores complejos, ya que a su vez son una estructura de rasgo. Algunos ejemplos de estructuras de rasgos se muestran en la tabla 2.5.

Palabra: **la**	Palabra: **amar**	Palabra: **hermosa**
<cat> = DET <conc num> = sing. <conc gen> = fem. <lex> = el	<cat> = V <arg0 cat> = SN <arg0 función> = sujeto <arg1 cat> = SN <arg1 función> = obj-dir	<cat> = ADJ <conc num> = sing. <conc gen> = fem. <lex> = hermoso

Tabla 2.5. Ejemplos de estructuras de rasgos

La información contenida en una estructura de rasgos se combina en una estructura nueva mediante la operación de unificación. Para que esto se produzca las estructuras deben tener información compatible, pues en caso contrario no se unificarían. La compatibilidad tiene que ver con la naturaleza de los rasgos y los valores. Los rasgos que solo aparecen en una de las estructuras unificadas se incorporaran a la estructura resultado de la unificación, logrando combinar la información común y diferente. Esto permite que diferentes estructuras informativas puedan ser combinadas coherentemente.

La tendencia de los lingüistas de utilizar gramáticas mas restringidas, los ha llevado a apreciar a las gramáticas independientes del contexto aumentadas con rasgos, a través del mecanismo de unificación [Shieber, S.].

2.3.1.2. Estructura de un sistema PLN simbólico

Por otra parte Moreno plantea que cualquier programa de NLP tiene 2 grandes tipos de conocimiento almacenado [Moreno, A.]:

1. *El conocimiento lingüístico*, en forma de gramática, léxico y modelo conceptual del mundo. La gramática es simplemente una definición abstracta de un conjunto de elementos estructurados y bien formados. Seria equivalente a nuestra competencia lingüística y pragmática.
2. *Programa o parser*, que contiene las instrucciones para procesar los datos lingüísticos. El parser es un algoritmo o conjunto de instrucciones que relacionan cadenas de símbolos con el conocimiento lingüístico almacenado.

El parser es un mecanismo computacional que infiere la estructura de las cadenas de palabras a partir del conocimiento almacenado en la gramática y diccionario, si establece si son cadenas gramaticales o agramaticales. El problema que tiene que resolver el parser es puramente sintáctico: reconocer las oraciones gramaticales y asignarles una estructura. Otros componentes se encargan de la interpretación. La siguiente figura muestra el esquema general del proceso de parsing [Winograd, T.] [Moreno, A.].

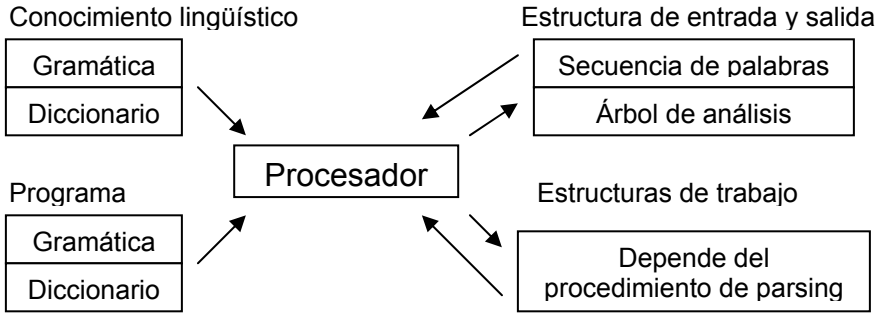

Figura 2.2. Esquema general del procesamiento de parsing (adaptado de Winograd, 1983) [Moreno, A.]

Los algoritmos de parsing son los encargados de decidir que reglas probar y en que orden. Cada algoritmo suele combinar diferentes parámetros y diferentes estructuras de trabajo. Hay muchos algoritmos, pero todos se basan en la combinación de tres parámetros esenciales que deben considerarse: análisis descendentes (top-down) / análisis ascendentes (bottom-up), procesamiento secuencial / procesamiento en paralelo, procesamiento determinista / procesamiento no determinista [Moreno, A.]. Considerando estos parámetros se

pueden mencionar algunos ejemplos de algoritmos de parsing: (1) Algoritmo descendente en serie con backtracking [Grishman, R.]. (2) Algoritmos con Chart [Winograd, T.] [Allen, J.]

Los modelos simbólicos son el paradigma predominante en CL, su repertorio de conceptos y métodos es muy amplio y ha sido aplicado sobre múltiples problemas y lenguas. De ellos los mas usados son los autómatas de estados finitos (por su sencillez y eficiencia de procesamiento) y las gramáticas independientes del contexto complementadas con gramáticas de unificación y rasgos (por su poder expresivo para dar cuenta de fenómenos lingüísticos) [Moreno, A.].

2.3.1.3. Ejemplo de modelos simbólicos: una gramática para el español

Una gramática independiente del contexto para algunas oraciones en español se muestra en la tabla 2.7. Esta gramática reconoce oraciones como “Cyrano ama a Melibea”, “el habla”, “ella ama a Cyrano” como gramaticales. Esta gramática al mismo tiempo genera la oración “el ama a el Melibea”, la cual no tiene una concordancia de géneros adecuada, por tanto no es gramaticalmente correcta aunque esta gramática la reconozca.

Reglas (Gramática)	**Diccionario** (Lexicón)
s --> ns, vs.	det --> [el].
s --> vs.	det --> [la].
vs --> v.	pron --> [el].
vs --> v, ns.	pron --> [ella].
vs --> v, ps.	n --> [cyrano].
vs --> v, ps.	n --> [melibea].
ns --> pron.	v --> [ama].
ns --> n.	v --> [habla].
ns --> det, n.	prep --> [de].
ps --> prep, ns.	prep --> [a].

Tabla 2.7. Una gramática independiente del contexto.

2.3.2. Modelos Estadísticos

La aplicación de la probabilidad y la estadística al estudio del lenguaje tiene una tradición al menos tan antigua como la de los modelos formales. La idea general es inferir conocimiento directamente de los datos, buscando regularidades significativas. Aplicando la estrategia general de contar con la mayor cantidad posible de datos para poder establecer una probabilidad lo más cercana posible a la frecuencia relativa estable.

En los años cincuenta Chomsky critica duramente estas teorías empiristas, proponiendo en su lugar las gramáticas formales. Por esta razón la lingüística

estadísticas ha sido muy minoritaria debido a la fuerza de las corrientes formales y cualitativas. Según Charniak, el estancamiento de los sistemas basados en el conocimiento se debe a que en ellos se asume que la comprensión de las lenguas naturales depende básicamente de una gran cantidad de "conocimiento del mundo", por lo tanto los sistemas de NLP tienen que contar con dicho conocimiento para tener éxito en su simulación de la facultad lingüística. Parece un hecho indiscutible el que, a pesar de los múltiples intentos realizados en IA, no se dispone de un modelo o marco formal para representar con éxito dicho conocimiento de "sentido común" que todo hablante parece emplear para entender los mensajes que recibe [Charniak, E.].

Además de la teoría de las probabilidades y la estadística, la Teoría de la Información [Lyons, J.] es uno de los fundamentos teóricos de los modelos estadísticos. La teoría de la información tiene por objetivo descubrir las leyes matemáticas que gobiernan los sistemas diseñados para comunicar y manipular información. Lyons señala los conceptos de rendimiento funcional (depende de la frecuencia de aparición de un contraste[12]) y contenido informático (depende de la probabilidad de aparición de una unidad en un contraste determinado) como los conceptos básicos en la teoría de la información.

Los modelos estadísticos, también llamados métodos cuantitativos, proporcionan una solución al gran problema de los modelos simbólicos: la ambigüedad. Cuando una oración presenta varias estructuras o interpretaciones posibles para escoger, se elegirá la más probable en función de las probabilidades de cada opción.

2.3.2.1. Modelos estadísticos en CL

Los modelos estadísticos más difundidos en NLP, mantienen los conceptos de teoría de la probabilidad (probabilidad condicionada[13] e independencia de sucesos).

Técnicas básicas, estimación y evaluación de probabilidades.

Las técnicas básicas consisten en calcular las frecuencias de las palabras que aparecen en un conjunto de textos, y deducir todas las probabilidades medias[14] y condicionadas. Aunque el uso de la probabilidad tiene escasa utilidad en la vida real; lo que se quiere es predecir acontecimientos a partir de cierta información incompleta, por ejemplo el análisis mas probable de una oración en un texto a partir de análisis anteriores.

---

[12] Contraste paradigmático: son dos o más elementos que pueden aparecer en la misma posición, en cualquier nivel lingüístico. Por ejemplo "r" y "rr" en caro y carro.

[13] Probabilidad condicionada: representa la probabilidad de una unidad con respecto a otra en un contexto concreto. P(A|B) probabilidad de A, dado B.

[14] Probabilidades medias: calcula de aparición de las unidades y su proporción sobre la muestra. Representa la estimación de la probabilidad de aparición de cada unidad de manera aislada.

Moreno nos reseña el método de estimación mas sencillo, que emplea frecuencias relativas extraídas de un conjunto de datos, llamado corpus lingüístico. El método tiene las siguientes etapas [Moreno, A.]:

1. Recolección de datos: obtener una muestra significativa de varios textos.
2. Anotación de las unidades del corpus[15]: el texto debe estar marcado con información para inferir estadísticas mas útiles. Las marcas mas habituales son morfosintácticas, para cada unidad se especifica su categoría, concordancia, etc. Aunque se puede anotar el texto con cualquier tipo de información pertinente, por ejemplo sintagmática o semántica-léxica. Un ejemplo de un corpus anotado con información sintáctica se muestra en la figura 2.4. La anotación puede hacerse manual o automáticamente. En la manera manual por lo general los primeros datos son marcados por especialistas que definan las ambigüedades, y luego se usan estos datos para el entrenamiento de anotadores automáticos. El método automático usa una gramática computacional para anotar el corpus. Existen algunas técnicas mixtas que requieren primero un tratamiento automático y después una revisión y corrección, el método mas utilizado actualmente es el Penn Tree Bank. Un ejemplo del texto anotado de un corpus en ingles generado por Penn Tree Bank es el siguiente:

```
( (S
  (NP Liberty/NP National/NP)
  (VP exchanged/VBD
   (NP (PP about/IN)
     78.64/CD shares/NNS
     (PP of/IN
      (NP its/PP$ common/JJ stock/NN)))
```

3. Calculo de frecuencias de las unidades: se calcula cuantas veces aparece cada unidad (palabra, sintagma, concepto, morfema, fonema, etc.) en función de que se hayan definido antes. Una posibilidad es calcular la probabilidad asociada a cada palabra con una determinada categoría sintáctica. Para ello se debe aplicar alguna técnica estadística (probabilidad condicionada, Ley de Bayes, n-gramas, árboles de decisión).

Para ilustrar este método básico se mostrara un pequeño ejemplo, para esto se dispone de un corpus con 10.255 palabras. Anteriormente se han etiquetado las palabras con sus categorías sintácticas, se calcularon sus ocurrencias y resultados parciales, se muestra en la siguiente tabla:

[15] Corpus: es gran conjunto de textos en lenguaje natural que incluye información extra tales como etiquetas para cada palabra indicando el constituyente gramaticales. Además por cada oración se forma un árbol de parse.

Palabra	Ocurrencias	Probabilidad
sobre, PREP	115 veces	115/10.255 = 0.0112
sobre, N	13 veces	13/10.255 = 0.0012
sobre, V	11 veces	11/10.255 = 0.0010
sobre	139 veces	139/10.255 = 0.0135

Tabla 2.8. Ejemplo de análisis de probabilidad de un corpus.

Con los cálculos anteriores se ha obtenido la probabilidad de ocurrencia de la palabra "sobre" con relación a toda la muestra. Por otra parte se calcula la probabilidad condicionada, con el fin de pronosticar cual categoría sintáctica (PREP, N o V) es mas probable que aparezca si la palabra es "sobre". Estos cálculos se muestran en la siguiente tabla:

Categoría	Probabilidad Condicionada P(A|B) = P(A&B) / P(B)
P(PREP|sobre)	0.0112 / 0.0135 = 0,829
P(N|sobre)	0.0012 / 0.0135 = 0,088
P(V|sobre)	0.0010 / 0.0135 = 0,074

Tabla 2.9. Ejemplo de análisis de probabilidad condicionada de un corpus.

Con estos cálculos se puede predecir que la palabra "sobre" tiene mas del 80% de posibilidades de ser una preposición.

Otra cuestión importante en estas técnicas básicas es decidir el tamaño de la muestra para obtener estimaciones fiables, con un margen de error aceptable. Con este propósito Kubácek en 1994 [Kubácek, L.] propone una formula para estimar el tamaño de una muestra representativa. Esta formula permite calcular N, el tamaño de la muestra, en función de dos variables: el numero de unidades consideradas (por ejemplo fonemas y categorías sintácticas) y la desviación estándar media de las frecuencias relativas.

Para finalizar hay que considerar el problema de los datos con muy baja frecuencia o incluso que no hayan aparecido nunca en el corpus (para el modelo su probabilidad es cero). Se origina entonces un serio problema de fiabilidad de estimación. Para solucionar esto hay que recurrir a métodos mas sofisticados, como por ejemplo la técnica estadística de n-gramas.

Modelo de N-gramas.

El modelo de N-gramas no es el mas empleado, pero si es el que mejores resultados ha obtenido. Debido a que permite mejorar la fiabilidad de la estimación teniendo en cuenta parte del contexto local. El modelo asume que solo unas pocas unidades anteriores condicionan la probabilidad de aparición de la siguiente unidad. Considerando a "n" como el numero de unidades que se tienen en cuenta,

el modelo obtenido se denomina n-grama. Los valores de "n" para casi todos los sistemas varían entre 2 y 7. El modelo mas común es el trigrama, donde se calcula la probabilidad condicionada de una unidad dada dos unidades precedentes [Charniak, E.].

Para crear un trigrama hay que utilizar un corpus de entrenamiento y registrar cada una de las parejas y tríos de palabras (o cualquier otra unidad) que aparezca en el texto. Los signos de puntuación se consideran también palabras. Se debe realizar el calculo de probabilidad de aparición de cada trío o pareja. Un ejemplo de un trigrama en base a categorías sintácticas de un fragmento de un corpus se muestra en la tabla 2.10. El calculo de la probabilidad de cada trío se realiza en base a la siguiente formula:

$P(U_i \mid U_{i-2}\ U_{i-1})$.

Los n-gramas basados en palabras suelen usarse en ciertas aplicaciones, ya que el dominio esta limitado al léxico usado en el corpus de entrenamiento. Con el fin de utilizarse en aplicaciones mas generales los n-gramas se calculan en función de muchos tipos de unidades, típicamente categorías sintácticas y fonemas (es el caso de los etiquetadores morfosintácticos y reconocedores del habla respectivamente). Los modelos estadísticos obtenidos sobre estos conjuntos de unidades son mas reducidos, pues manejan etiquetas de pocas decenas de categorías sintácticas, en lugar de los varios millones de palabras que consideran los n-gramas para palabras.

Si las probabilidades estimadas por este método se añaden a un autómata de estados finitos, entonces obtenemos un autómata probabilístico, también conocido como cadenas de Markov[16] [Charniak, E.]. Los modelos markovianos asumen que las gramáticas de las lenguas naturales son de estados finitos. Estos tiene gran aplicación practica en sistemas de reconocimiento de habla, básicamente por su eficiencia. Esta eficiencia se debe a que las restricciones locales de algunas lenguas naturales son muy fuertes, según las investigaciones de [Church y Mercer]. Las unidades lingüísticas parecen bastante condicionadas por el contexto, especialmente las restricciones del tipo semántica-léxico.

Parejas	Probabilidad	Tríos	Probabilidad
ART N	0,6	ART N V	0,4
N V	0,7	N V N	0,2
N ART	0,2	V ART N	0,5
V ART	0,8	N V ART	0,6

[16] Cadena de Markov: es una red que utiliza probabilidades. Es un tipo de proceso estocástico, es decir, un conjunto finito de variables aleatorias encadenadas que tienen una probabilidad conjunta. La suma de las probabilidades de los arcos debe ser 1.

V N	0,4		

Tabla 2.10. Trigrama para categorías de un corpus de ejemplo

La técnica de los n-gramas tiene varios problemas. Uno de ellos es el tratamiento de los datos no encontrados en el corpus de entrenamiento. Pues siempre habrá alguna combinación nueva en el conjunto de prueba para la cual no se tiene calculada alguna probabilidad o la probabilidad es cero. Una técnica habitual es suavizar las probabilidades mediante un calculo adicional de bigramas y unigramas [Allen, J.], utilizando estos cálculos cuando aparezca un trigrama con probabilidad cero, se obtendrá alguna estimación del bigrama y del unigrama.

Existen muchos modelos estadísticos, se han presentado la evaluación y calculo de probabilidades y los n-gramas, considerados los mas básicos y utilizados. Otros modelos estadísticos son el Canal de Ruido usado para reconocimiento de habla y cadenas de Markov ocultas empleadas para desarrollar anotadores estocásticos de textos.

También se han realizado aplicaciones en donde se pueden implementar gramáticas independientes del contexto, donde se tengan estadísticas del uso de cada regla gramatical. Además hay serios proyectos en construir una gramática computacional a partir del uso de corpus y estadísticas [Black, Garside y Leech].

Dentro de las limitaciones de los modelos estadísticos podemos destacar varios. La representatividad del corpus es probablemente el problema mas importante de todo modelo estadístico en general: pues son totalmente dependientes del corpus, de manera que si se intenta aplicar el modelo a otro dominio los resultados son pobres. Otra limitación tiene que ver con la localidad: es muy eficiente con las relaciones locales, pero incapaz con las relaciones a larga distancia, mientras las gramáticas sintagmáticas tiene medios para tratar con elementos discontinuos, la estadística parece que no ha encontrado soluciones [Lyons, J.].

2.3.2.2. Ejemplo de modelos estadísticos: una gramática probabilística para el español

Como se explico antes, el estudio probabilstico de un corpus, requiere recoger datos sobre el empleo de cada constituyente y secuencia de estructuras gramaticales. No solo se pueden determinar secuencias de pares y tríos como los trigramas. También se puede recoger la estadística del uso de una regla gramatical. Con esta idea se forman gramáticas sintagmáticas probabilísticas, como se puede observar a continuación.

Se asume que se tiene un corpus de entrenamiento como el siguiente:

[S [SV [V Fuimos] [SP [P al] [SN [N cine]]]]]
[S [SV [SN [PRO Me]] [V gusta]] [SN [DET el] [N pescado]]]
[S [SN [N Juan]] [SV [V colecciona] [SN [N sellos]]]]
...
[S [SN [V Vi] [SP [P a] [SN [DET un] [N hombre] [SP [P con] [SN [DET un] [N telescopio]]]]]]

Figura 2.4. Muestra parcial de un Corpus de entrenamiento.

Se puede determinar la cantidad de veces que se aplica una regla gramatical y realizar los cálculos estadísticos anteriormente explicados. La siguiente tabla muestra este resultado.

Regla	**Ocurrencia de α**	**Ocurrencia de la Regla $\alpha \rightarrow \beta$**	**Probabilidad**
S → SN SV	10	5	0.5
S → SV	10	3	0.3
S → SV SN	10	2	0.2
SV → V SN	10	1	0.1
SV → V SP	10	4	0.4
SV → SN V	10	1	0.1
SV → V ADJ	10	1	0.1
…	…	…	…
SN → N	17	8	0.47
SN → DET N	17	1	0.06
SN → DET N SP	17	1	0.06
SP → P SN	6	6	1

Tabla 2.11. Muestra parcial de una gramática probabilística.

Con esta gramática se puede calcular la probabilidad de cada uno de los análisis posibles de una oración en particular. Si una oración tiene mas de un posible análisis sintáctico entonces se resolverá la ambigüedad con los valores de probabilidad de aplicar las reglas que están en conflicto. La regla que tenga mayor probabilidad será escogida.

2.3.3. Modelos Biológico.

Cualquier sistema NLP esta organizado en diferentes módulos (reconocimiento léxico y morfológico, análisis sintáctico, interpretación semántica y pragmática). La mayoría de estos sistemas usan una estrategia lineal, que no se corresponde con el procesamiento simultaneo y en paralelo que realiza nuestro cerebro. Si al procesar una emisión lingüística también consultamos simultáneamente diferentes componentes lingüísticos, entonces las aproximaciones inspiradas en el funcionamiento neuronal puede ser una manera de acercarse no solo a una

simulación del cerebro sino a un procesamiento del lenguaje más eficiente [Moreno, A.]. Con estas ideas surgen entonces los modelos biológicos aplicados al PLN.

Se dividen en dos grandes grupos: los inspirados en el cerebro, el conexionismo (redes neuronales) y los inspirados en la vida y la evolución, la computación evolutiva (algoritmo genético).

En palabras de los dos conexionistas mas destacados, Rumelhart y MacCleland, el comportamiento de un sistema cognitivo no esta gobernado por reglas, sino más bien descrito por reglas de manera aproximada [Rumelhart y MacCleland]. Según los conexionistas, el aprendizaje y el conocimiento operan en un nivel por debajo de las reglas simbólicas clásicas, es decir, a un nivel subconceptual y subsimbólico. Las redes conexionistas utilizan habitualmente cálculos estadísticos para la asignación de los pesos, por tanto se asemejan a los modelos probabilísticos.

2.3.3.1. Redes Neuronales (Conexionismo)

Las redes neuronales imitan el comportamiento del órgano cerebral. En un modelo conexionista no están explícitos ni una gramática, ni un lexicón. En su lugar existe un reconocimiento de estructuras (fonemas, morfemas, oraciones) que se realiza sobre la base de semejanzas en los patrones de activación de nodos: dos estructuras son similares si excitan los mismos nodos.

Una de las características mas destacadas de la redes neuronales es el procesamiento distribuido en paralelo: los diferentes procesos de reconocimiento de palabras, análisis morfosintáctico y semántica actúan simultáneamente. Desde la perspectiva conexionista el procesamiento consiste en la acción conjunta de múltiples unidades distribuidas actuando en paralelo. Debido a la redundancia y al carácter aproximado del procesamiento, estos sistemas intentan explicar por que los hablantes son capaces de superar las deficiencias y los errores en la comunicación lingüística: aunque alguna unidad implicada en el procesamiento falle el mensaje suele interpretarse gracias a que el conjunto de unidades es muy numeroso.

Si bien los modelos conexionista suponen un planteamiento radicalmente diferente a los modelos simbólicos, no usan símbolos, la oposición de este ante el modelo estadístico no es tan radical. Esto se debe a que generalmente las redes conexionista realizan cálculos estadísticos para la asignación de pesos.

Algunas de las aplicaciones en donde los modelos conexionistas se han aplicado en el PLN, debido a que implican un reconocimiento de patrones, son las siguientes:

1. Reconocimiento del habla: implica el procesamiento de grandes cantidades de datos acústicos, los cuales pueden organizarse en niveles (rasgos acústicos, alófonos, fonemas, sílabas y palabras).
2. Desambiguación léxica: son útiles para establecer relaciones entre palabras, simulando un lexicón mental. Empleado para resolver ambigüedades
3. Etiquetadores morfosintácticos: reconocimiento de signos de puntuación y palabras homógrafas.

La capacidad de aprender de estos sistemas, es una de las características mas apreciadas. Esta capacidad de aprendizaje permite una mayor adaptación de estos sistemas a un nuevo dominio, una vez construida la red neuronal, debe realizarse un entrenamiento con ejemplos del nuevo dominio.

Entre las limitaciones de los modelos conexionistas podemos señalar que, hasta la fecha, han tenido poco éxito en el procesamiento sintáctico y semántico de oraciones. Esto se debe a su incapacidad para tratar la recursión [Uszkoreit, H.], la cual es considerada uno de los rasgos mas destacados de las lenguas naturales.

2.3.3.2. La computación evolutiva: Algoritmos genéticos

El padre de los algoritmos genéticos es John Holland, en 1975 publico un clásico de AI titulado “Adaptación en sistemas naturales y artificiales”, donde se expuso por primera vez un paradigma computacional que imita a los sistemas complejos adaptativos. A partir de los años ochenta, un área de investigación multidisciplinaria denominada Teoría de la Complejidad, contribuye con el desarrollo de los algoritmos genéticos [Waldrop, M.]. Esta estudia los agentes independientes que interactúan entre sí de diferentes maneras. Además este paradigma se caracteriza por la auto-organización espontánea (los elementos buscan equilibrio) y por ser adaptables (aprender de la experiencia y adaptarse al entorno).

Los algoritmos genéticos es una de las técnicas mas utilizadas en la simulación de los sistemas complejos adaptativos. La computación evolutiva se basa en la capacidad de la naturaleza de resolver problemas. La idea central en esta técnica es reproducir el entorno en el que se produce una evolución. Los sistemas están compuestos por individuos que interactúan estableciendo relaciones muy complejas de competencia y colaboración. Una parte importante de los sistemas de un modelo de simulación evolutiva son los mecanismos de evaluación, selección y reproducción. Se debe escoger según el problema la técnica o estrategia que mejor se adapte.

Las lenguas naturales pueden ser consideradas un sistema complejo adaptativo. Estos sistemas se caracterizan por su dinamismo y evolución. El lenguaje natural es suficientemente estables como para ser aprendido y permitir la comunicación entre los miembros de la comunidad lingüística. Al mimos tiempo son lo

suficientemente inestables como para contar con variantes (por ejemplo dialectos) de las cuales unas desaparecen y otras sobreviven [Moreno, A.].

Según Moreno, cualquier nivel lingüístico puede someterse a experimentación en una simulación evolutiva. Sin embargo el nivel fonológico y léxico son los que mejor se prestan para esta representación [Moreno, A.].

2.3.3.3. Ejemplo de modelos biológicos: tratamiento de fonemas en español

El nivel fonológico puede ser simulado con un algoritmo genético de una manera , debido a que los cambios fonéticos se producen en base a un sistema de reglas de la gramática. Además estos cambios son mas simples, concretos y cuantificables que los otros cambios lingüísticos.

En este ejemplo de Moreno se expondrá la manera de codificar las unidades (fonemas) y sus mutaciones [Moreno, A.]. La forma habitual de codificar las unidades es mediante cadenas de dígitos binarios. Los fonemas tienen rasgos distintivos y en base a ellos se puede obtener una codificación binaria. Un ejemplo simplificado, se puede observar en la tabla 2.12. en donde /b/ es representado como la cadena 11001###.

	/p/	/b/	/m/	/t/	/d/	/n/	/a/	/i/	/u/
Consonante	1	1	1	1	1	1	0	0	0
Labial	1	1	1	0	0	0	#	#	#
Dental	0	0	0	1	1	1	#	#	#
Nasal	0	0	1	0	0	1	#	#	#
Sonora	0	1	1	0	1	1	1	1	1
Anterior	#	#	#	#	#	#	0	1	0
Central	#	#	#	#	#	#	1	0	0
Posterior	#	#	#	#	#	#	0	0	1

Tabla 2.12. Fonemas en español en formato binario. (1 significa un valor positivo, 0 un valor negativo y # significa valor no pertinente).

Los procesos de reproducción y mutación generalmente trabajan con codificaciones binarias. Las reglas se mutación se aplican según la zona de la cadena binaria, en donde la regla cambia un valor concreto. Por ejemplo un proceso de mutación que convierta una consonante de oclusiva a nasal, solo cambia el valor del cuarto dígito de la cadena.

Lo que se pretende con este ejemplo, es explorar hipótesis evolutivas de ciertos fenómenos y factores que influyen en el conjunto de fonemas. Lo cual es de gran importancia para los lingüistas teóricos.

2.3.4. Comparación de los modelos de Procesamiento de Lenguaje Natural

Los tres modelos de NLP fueron explicados anteriormente, se presenta a continuación un cuadro comparativo:

	Modelo simbólico	**Modelo estocástico**	**Modelo biológico**
Teorías	Teoría de conjuntos y lógica matemática	Teoría de la información y Estadística.	IA, Conexionismo, Teoría de la Complejidad.
Defensores	Chomsky, Minsky	Shannon, Skinner, Harris.	Rumelhart, MacCleland, Holland.
Idea	Representar la estructura lógica del lenguaje (conocimiento del lenguaje). Conocimiento = la competencia.	Inferir conocimiento lingüístico a partir de los datos (corpus). Conocimiento = la actuación	Simular la capacidad lingüística, el aprendizaje y la evolución del lenguaje natural. Conocimiento = la actuación y la competencia.
Consiste	Sistemas formales axiomáticos, formados por reglas y símbolos. Propiedades de elementos y relaciones.	Sistemas estocásticos, sus probabilidades se determina del estudio de un contexto. Datos o corpus, cálculos estadísticos, reglas lingüísticas + probabilidades.	Sistemas que simulan la naturaleza, aplican técnicas de aprendizaje y representaciones simbólicas que evolucionan. Elementos, reglas, entrenamiento y pruebas
Técnicas	Gramáticas: generativas, de estados finitos, independientes del Contexto, Unificación y Rasgos.	N-gramas Cadenas de Markov Árbol de Decisión Gramáticas probabilísticas	Algoritmos genéticos Redes neuronales
Ventajas	Están definidos formalmente Facilitan la evaluación de hipótesis Control del conocimiento	Mayor eficiencia Resolución de ambigüedad por cálculos de probabilidad Conocimiento de uso y aceptabilidad del lenguaje	Adaptación directa a nuevos dominios (entrenamiento) Procesamiento en paralelo Toleran los errores humanos en la comunicación lingüística
Limitaciones	La ambigüedad Incorporar semántica requiere heurísticas Cambios de contextos requieren gran esfuerzo	Dependen del corpus de entrenamiento. Relaciones de localidad No aplican conocimiento sobre el lenguaje	Hay poco desarrollo conceptual y técnico
Aplicaciones	Comprensión Correctores ortográficos Correctores sintácticos y de estilo	Reconocimiento Etiquetadores estocásticos Desambiguación léxica y sintáctica Reconocimiento del habla Traducción automática	Reconocimiento del habla Simulación y Reconstrucción de lenguas Modelos de evolución de lenguas Planificación lingüística

2.4. Revisión Histórica del NLP

Moreno afirma que el desarrollo histórico de los sistemas de NLP ha estado marcado por desarrollos teóricos y técnicos en Lingüística y en Computación. En las ultimas décadas la CL ha estimulado la aparición de nuevas ideas en los mencionados campos y se ha convertido en un importante evaluador de teorías.

Así presenta una breve reseña de la historia del NLP, desde los años 50 hasta finales de los 90 [Moreno, A.].

Cincuenta y Sesenta: *Inicio, primeros problemas*

Los trabajos pioneros en el campo del NLP se dieron en los años cincuenta y principios de los sesenta y, concretamente en traducción automática [Locke y Booth]. Estos sistemas fueron un fracaso por las siguientes razones: bajo nivel de los conocimientos del lenguaje y de la lingüística matemática, y la baja capacidad de procesamiento de las computadoras de la época. Además no existían lenguajes de programación para trabajar eficientemente con palabras y símbolos.
La investigación lingüística se enfoca al campo de la lingüística estadística (estudio del léxico de autores, confección de diccionarios de frecuencias, elaboración de índices y concordancia). Los primeros trabajos en traducción automática asumían muy poco o ningún conocimiento lingüístico teórico.
A finales de los cincuenta, Chomsky introdujo sus ideas revolucionarias. Fue el primero en introducir el paradigma lógico formal caracterizado por unidades y reglas, así como también la noción de recursividad, composición y creatividad de las lenguas naturales. Propuso una jerarquía de gramáticas generativas, lo cual representa un punto de referencia dentro de la teoría de lenguas formales y autómatas [Chomsky, N.].
En los años sesenta hubo un desarrollo aceptable de las Gramática Generativas. Así en 1962, se fundó la Association for Computational Linguistics, ACL, con el fin de desarrollar técnicas, métodos y aplicaciones, estableciendo limites con otras áreas del conocimiento.
En 1966, un famoso informe de la National Academy of Sciences de los EEUU afirmaba que con la tecnología existente en la época no se podía alcanzar ningún éxito en traducción automática. En este mismo año hay una evidencia substancial de un generador de índices automático denonimado el proyecto SMART [Salton, G.,]. Usa las técnicas de análisis de texto automático en recuperación de información de esta década (agrega índices ponderados, cada termino del índice es un concepto, uso de varios diccionarios).
En esta década se pusieron de moda los programas que creaban poesía al azar. A pesar de que los programas en si mismos son bastante triviales, en su momento causaron sensación. La estructura de estos se basaba en la creación de patrones, creación de una BD y la utilización de una función aleatoria [Moreno, A.].

Setenta: *primeros sistemas funcionales*

El primer sistema funcional se produjo con la aparición del programa SHRDLU de Winograd, en 1971 [Winograd, T.] , marcando la diferencia con los primeros sistema de NLP. Demostrando que es posible que un computador entendiera una lengua natural en un dominio restringido. Este podía interpretar preguntas y ordenes sencillas, así como realizar inferencias, explicar sus acciones y aprender nuevas palabras.

En los sistemas de los años setenta la gramática y el parser estaban entremezclados dentro del programa. Las técnicas mas extendidas para escribir gramáticas computacionales fueron las Redes de Transición Recursiva (RTN) y sus derivadas, las Redes de Transición Aumentadas (ATN). Una red de transición de estados es una representación de una gramática regular o de estados finitos. Las redes de transición recursivas permiten representar estructuras recursivas en forma de subredes. Las redes de transición aumentadas se desarrollan para tratar problemas típicos de las gramáticas transformacionales, son en realidad un lenguaje de programación para construir analizadores sintácticos [Gazdar y Mellish]. Los programas construidos con estas técnicas son rígidos y difíciles de reutilizar [Moreno, A.].
A finales de los setenta, luego de un largo periodo de inactividad investigativa, se retoman los proyectos en traducción automática, gracias a los considerables avances de experimentados en lingüística y computación.
Algunos ejemplos de los proyectos de esta epoca son:

- ELIZA, que imitaba a un psiquiatra [Covington, M.];
- MEDLINE, un sistema de extracción de información sofisticado en línea [McCARN, D.B.];
- LADDER, sistema que en 1978 se valió de una gramática semántica para tener acceso a bases de datos de la Marina de los Estados Unidos [Hendrix, G.];
- INTELLECT de Harris, que ahora es un producto comercial, también se basa en reglas gramaticales [Harris, L.];
- Rendezvous de Codd, sistema que usaba un léxico y convertía consultas expresadas en lenguaje natural a expresiones del calculo relacional con base en una gramática de frases [Codd, E.].

Ochenta: *Lenguajes declarativos y gramáticas no transformacionales*

En los años ochenta hubo un cambio radical en las técnicas utilizadas en los sistemas de NLP. Gazdar y Mellish en 1989 señalan que la mayoría de los avances en NLP en los setenta y ochenta se debió a un cambio en el enfoque teórico y practico de la informática [Gazdar y Mellish]. En el plano lingüístico, los investigadores empezaron a explorar las ventajas de utilizar formalismos gramaticales más sencillos que las gramáticas transformacionales. En el plano informático, el estilo declarativo se fue imponiendo. Los sistemas se van haciendo más flexibles, y portables, pues los sistemas de estilo procedural exigían diferentes gramáticas para generación y análisis. Todo esto se consiguió gracias a tres importantes innovaciones en esta década: los formalismos gramaticales de unificación [Shieber, S.], los lenguajes declarativos de programación lógica como Prolog, y los chart parsers [Moreno, A.].
Los formalismos de unificación permiten definir gramáticas independientes del contexto con información sintáctica, con la idea básica de que la gramática contenga reglas sencillas y que sean la información léxica y la unificación las que lleven todo el peso del procesamiento. Se crearon entornos de programación para

desarrollo de gramáticas de unificación, como PATR, y Shieber es precisamente uno de sus creadores, se trata de un lenguaje de programación que permite codificar información lingüística.
Con el uso de Prolog se libera al lingüista de pensar en los problemas de procesamiento, debido a que es declarativo y permite realizar directamente la unificación. Los chart parser, es una técnica de desarrollo de analizadores sintácticos que almacena resultados intermedio durante el procesamiento para mejorar la eficiencia, la robustez y los tiempos de respuesta.
A principios de los ochenta surgió un nuevo tipo de gramáticas generativas (uso de rasgos y unificación). Su utilización comenzó con el modelo de M. Kay (Funtional Unification Grammar), esto se extendió a teorías lingüísticas como LFG (Lexical Functional Grammar), GPSG (Generalized Phrase Structure Grammar) o HPSG (Head-Driven Phrase Structure Grammar) [Pollard y Sag]. Estas teorías han inspirado diversas gramáticas computacionales.
Los primeros sistemas que aplicaban modelos probabilísticos tenían como objetivo estudiar la variación lingüística, con el fin de determinar regularidades estadísticas.
A partir de los años ochenta se produce un cambio de tendencia, se pretende modelar el lenguaje natural con modelos puramente estadísticos. Los grupos pionero en esta tarea se dieron en la costa este americana (Pennsylvania, AT&T Bell Labs e IBM Yorktown), teniendo influencia sobre los avances en el procesamiento del habla.

Los noventa: *ascenso de los modelos probabilísticos*

Los años noventa también han supuesto un cambio de tendencia. Los sistemas de los ochenta estaban basados fundamentalmente en el conocimiento gramatical, con una amplia cobertura sintáctica, y en la medida que estos sistemas se fueron haciendo más complejos, se hicieron evidentes los limites del conocimiento lingüístico. El conocimiento lingüístico de los años ochenta estaba basado en la competencia del lingüista (modelo teórico), y los sistemas de NLP son ante todos sistemas prácticos que tienen que resolver casos reales de uso, y que deben responder eficazmente a problemas concretos. Debido a esto en esta década surgen dos acciones: la búsqueda de aplicaciones mas realistas (sistemas de ayuda y sistemas asistidos), y la ampliación del sistema a cualquier tipo de texto (sistemas con dominios no restringidos, capaces de discriminar la información relevante) [Moreno, A.].
En la actualidad existen entornos de desarrollo de sistemas NLP que facilitan enormemente el trabajo de los lingüistas computacionales. Algunos de ellos como la plataforma ALEP, desarrollada por la contribución de la Unión Europea.
En esta década hay un claro giro hacia una parte mas aplicada y comercial, favoreciendo el resurgir de las técnicas probabilísticas basada en corpus de datos Lingüísticos. En la actualidad, muchas aplicaciones mezclan conocimiento declarativo con conocimiento estadístico para mejorar las limitaciones inherentes a cada modelo, sobre todo para resolver el problema de la ambigüedad [Moreno, A.].

3. Definición del Problema

Cuando las computadoras de alto rendimiento llegaron a estar disponibles para el trabajo no numérico, se pensó que el computador podía "leer" una colección completa de documentos para extraer la información relevante. Pronto llegó a ser evidente que usando el texto en lenguaje natural de un documento no solamente se tenían los problemas del almacenaje, sino también resolver el problema de caracterizar el contenido del documento.

Pero la caracterización automática en la cual el software intenta duplicar el proceso humano de la "lectura" es un problema muy ambicioso. Específicamente, la "lectura" implica extraer la información, sintáctica y semántica, del texto y usar esto para extraer la información relevante del texto y descartar la no relevante. Al Aplicar técnicas de Recuperación de Información en los años 70, van Rijsbergen da una introducción sobre este tratamiento en donde pone de manifiesto la necesidad de un procesamiento del lenguaje natural [van Rijsbergen. C.J.].

Trabajos mucho más recientes como los de Lewis, nos indican que nuevas demandas en la recuperación de información proveen oportunidades al procesamiento del lenguaje natural para trabajar con técnicas ya probadas de recuperación estadísticas. La descripción compacta del contenido relevante de un documento puede incrementar la eficiencia de la clasificación del material textual como relevante y no relevante, pues justamente la selección de características es critica en las tareas de clasificación. Las investigaciones en el área de recuperación de texto tienen evidencias que sugieren que para las búsquedas de los usuarios finales, los índices de lenguaje deberían ser de lenguaje natural, en lugar de ser orientados al lenguaje controlado [Lewis y Sparck].

Esto indica que, el procesamiento del lenguaje natural parece ser un problema común de muchas aplicaciones que manejan información y conocimiento. Precisamente lo que se tiene como objetivo en este trabajo es desarrollar una herramienta que procese texto en español y extraiga sus descriptores o palabras claves. La idea es que se tenga una representación resumida del documento, para que estos descriptores puedan servir a otras aplicaciones o módulos que procesen grandes volúmenes de datos y textos en lenguaje natural. Estas aplicaciones pueden ser por ejemplo buscadores eficientes, herramientas de extracción de conocimiento, minería de texto, catalogadores automáticos, etc., que utilizaran la salida de nuestra herramienta como fuente de información.

El problema ahora es ¿cuál es la gramática más apropiada para describir formalmente las lenguas naturales? Moreno plantea que todo el mundo esta de acuerdo en que la respuesta debe conjugar 2 propiedades: (1) Expresividad: la gramática tiene que ser lo suficientemente poderosa como para abarcar todas las construcciones posibles en las lenguas naturales. (2) No sobre generación: la

gramática tiene que ser suficientemente restringida para no permitir como válidas construcciones agramaticales. Estos funciona en la teoría, pero la práctica las gramáticas formales se van modificando según las necesidades particulares: se introducen restricciones para reducir su poder o se incluyen extensiones para aumentar su expresividad [Moreno, A.]. Las gramáticas expresan la sintaxis y la morfología y estos son diferentes entre los distintos idiomas, por tanto una gramática pierde la generalidad pues esta atado a un léxico (vocabulario del dominio) y a un idioma.

Según Moreno cualquier gramática computacional en español de cierta cobertura basada en la competencia (conocimiento lingüístico) produce un numero muy grande de análisis sintácticos alternativos para la mayoría de las oraciones. Sin embargo, es significativo que los hablantes no noten tanta ambigüedad cuando procesan oraciones: de manera espontánea solo tiene 2 o 3 posibilidades [Moreno, A.]. Esto se debe a que la interpretación ayuda a desambiguar sintácticamente, ya que los hablantes prefieren las interpretaciones plausibles a las poco probables. La plausibilidad de una interpretación viene dado por el contexto semántico-pragmático. La solución para reducir el numero de análisis y la ambigüedad sintáctica es incorporar restricciones semánticas y pragmáticas muy finas. Estas restricciones significan una manera de incorporar el procesamiento del conocimiento del mundo en el sistema de NLP.

El conocimiento del mundo, es la representación del mundo real del hablante y de su dominio de conocimiento. Este conocimiento del mundo interviene en el nivel interpretativo del lenguaje o también llamado la semántica. Este nivel interpretativo es mas general que el nivel lingüístico, pues parece no depender directamente del idioma. Por ejemplo Williams plantea reglas de estilos y estas pueden ser aplicadas no solo al idioma ingles, para el cual fueron concebidas [Williams, J.]. Así se pueden tener varios nivel interpretativos del lenguaje: el primero es la semántica de las oraciones aislada. Esta semántica oracional pueden ser expresada de manera lógica, por ejemplo "Abelardo ama a Eloisa" se representa lógicamente como la cláusula ama(Abelardo, Eloisa). El segundo nivel es la semántica del discurso, relación de las oraciones entre si, la cual se puede simbolizar también a través de reglas lógicas.

El conocimiento del mundo se ha logrado expresar con técnicas de representación del conocimiento, tales como modelos, redes semánticas y ontologías. El conocimiento del mundo del hablante, se puede entender entonces como una base de conocimiento. Este conocimiento del mundo interviene en la interpretación oracional y del discurso, pues en base a dicho conocimiento se maneja el conocimiento implícito de los hablantes, que permiten resolver las ambigüedades, y también se resuelven las anáforas y elipsis. El nivel interpretativo se puede definir como un conjunto de reglas lógicas que procesan la información gramatical (sintáctica) y el conocimiento del mundo.

Entonces se puede pensar que un procesamiento lógico permitiría logra un nivel interpretativo en una gramática. Las reglas lógicas para la interpretación oracional y del discurso, pueden complementar las reglas gramaticales. El mismo nivel interpretativo, basado en reglas lógicas, podría procesar la representación del conocimiento del mundo que debe tenerse para el dominio de conocimiento.

Un modelo como el mostrado en la figura 3.1. ilustra el esquema de un sistema de procesamiento de lenguaje natural, en donde se tiene un modulo gramatical (nivel morfológico y sintáctico) y un modulo interpretativo (nivel semántico y pragmático). La base de conocimiento del mundo permite relacionar el conocimiento lingüístico con el contexto (conocimiento del mundo).

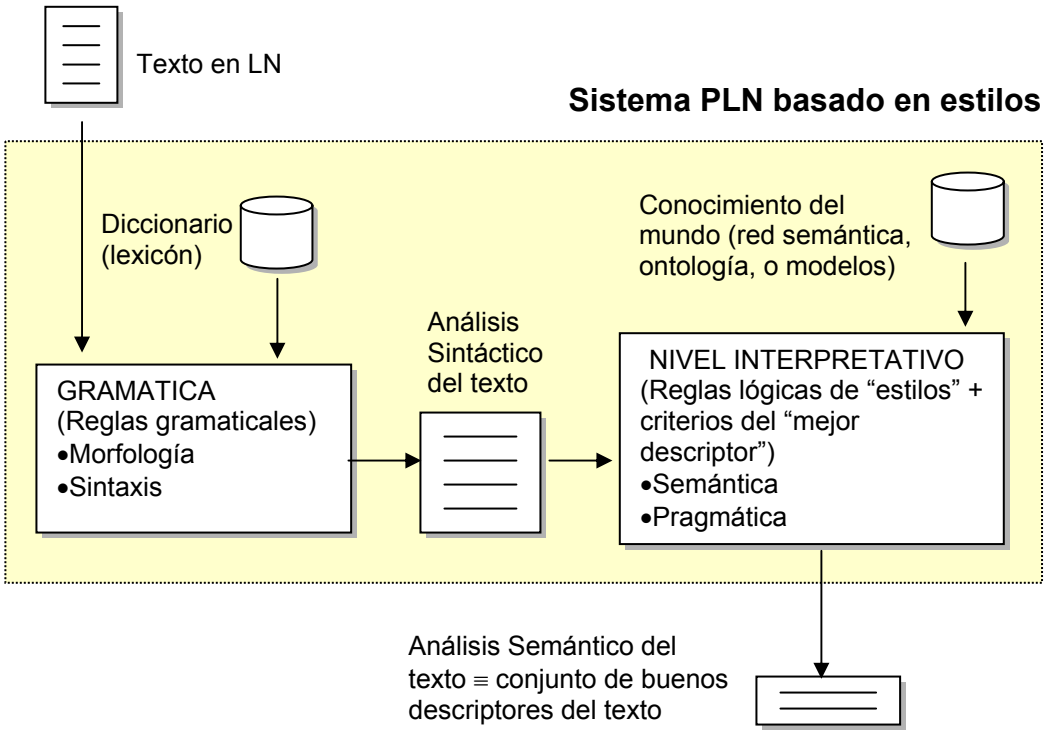

Figura 3.1. Un diagrama del Procesamiento del lenguaje natural con una "gramática de estilos".

Por otro lado en [Wiebe, Hirst y Horton], se plantea que cualquier texto establece un contexto lingüístico, sobre el cual las siguientes palabras deben ser entendidas. El escritor de un texto se dirige a un lector con el propósito de informar, divertir, colaborar en una tarea, quizás. Los lectores deben inferir la intención subyacente como parte de su comprensión. En el articulo de Wiebe se presentan recientes investigaciones en el uso del lenguaje en un contexto, y concluye que el uso del lenguaje involucra mucho mas que creación y comprensión de palabras aisladas, por tanto las computadoras, al igual que las personas, deben alojar el contexto interpersonal y lingüístico si están usando el lenguaje de una manera natural.

Precisamente un grupo de estas investigaciones en el uso del lenguaje en un contexto, tratan de expresar el matiz y el estilo en el lenguaje [Wiebe, Hirst y Horton]. La exacta escogencia de palabras, frases y estructura de las oraciones afectan el significado y el efecto preciso de una palabra. Un escritor elige (conscientemente o no) como objetivos por ejemplo ser formal o amigable, persuasivo o despectivo, claro u oscuro. Estos aspectos de una palabra son mucho más parte de su mensaje que su significado literal, y cualquier sistema de lenguaje natural sofisticado debe ser sensible a ello. Un problema particular son las expresiones referidas: palabras o frases que un escritor usa para denotar algún objeto o entidad particular. Mucho de los trabajos en la generación del lenguaje, incluyen las expresiones referidas, buscando determinar el contexto en el que son expresadas.

Según los trabajos de DiMarco y Hirst [DiMarco y Hirst], un escritor usa varias construcciones sintácticas con un objetivo estilístico de alto nivel. Con el fin de asegurar que una traducción automática retenga estos objetivos si ellos requieren una estructura sintáctica diferente en el lenguaje destino. Para capturar esta clase de intuición lingüística, estos investigadores desarrollaron la idea de una "gramática de estilos", la cual relaciona las estructuras sintácticas de un lenguaje con un conjunto de objetivos estilísticos independientes del lenguaje. En las tareas de traducción, este objetivo puede ser determinado en el texto origen y ser usado en la generación del nuevo texto.

Problema de la propuesta

Lo que se pretende con este proyecto entonces es diseñar una herramienta para asociar descriptores a textos en lenguaje natural en el idioma español. Esta herramienta complementa el análisis gramatical tradicional con reglas de estilo como las sugeridas por J. Williams en [Williams, J.] (ver figura 3.1.). En la propuesta de Williams, a cada oración en un texto se le asocia un tópico y la secuencia de estos tópicos en un párrafo sirve para analizar su coherencia. Nuestra intención es usar los tópicos como información básica para extraer descriptores significativos de una colección de párrafos en un texto. En la figura 3.1. se muestra como los estilos de Wiliams permitirá realizar un análisis interpretativo del texto. Este análisis además cuenta con criterios para escoger entre los tópicos los mas adecuados para describir el texto, lo cual se fundamenta también del conocimiento del dominio de información al cual pertenezca dicho texto.

El problema en concreto puede ser expuesto como sigue:

- Definir una estrategia para interpretar texto en español y extraer sus descriptores.
- Explorar criterios lógicos para asociar descriptores adecuados y relevantes a

textos en español.

Hipótesis:

- Si se aplica una gramática de estilos entonces se pueden obtener buenos descriptores de un documento escrito en base a estos estilos.
- Si la gramática esta extendida con reglas de estilos y el texto esta basado en las reglas de estilos entonces los descriptores que se derivan de la estrategia serán próximos a los descriptores obtenidos por los expertos en el dominio.

4. Metodología: ¿Cómo? ¿Cuál es la estrategia?

Los pasos a seguir para verificar las hipótesis planteadas en el problema consisten en los siguientes pasos:

(A) Obtener un Modelo de la Gramática basada en estilos:

Consiste en la definición formal de la gramática que use un procesamiento lingüístico simplificado (sintaxis) y modificado con reglas de estilo (semántica) para procesar textos y extraer buenos descriptores.

Lo que hasta ahora se tiene planteado es realizar una gramática sencilla basada en los estilos de Williams, para darle la semántica adecuada al texto y así identificar las acciones y los actores.

Esta gramática establecerá la semántica que permita identificar los descriptores, palabras claves o tópicos con los cuales obtener una descripción breve del texto.

(B) Establecer criterios para definir un "Buen Descriptor":

Como dice van Rijsbergen, C. J., es intelectual posible que un ser humano determine el contenido relevante de un documento. Para que un computador haga esto se necesita construir un modelo dentro del cual se construyan las decisiones de relevancia. Es interesante observar que la mayoría de la investigación en la recuperación de información muestran haber tratado diversos aspectos de tal modelo [van Rijsbergen. C.J.].

Un buen descriptor para nosotros sería una consecuencia lógica del discurso que estemos analizando

Discurso |= descriptor.

En este caso no interesa cualquier consecuencia lógica (pueden haber muchas), Se escogera la mejor:

rep-del-Discurso U Teoria-de-apoyo |= buen_descriptor(descriptor).

La teoría de apoyo es una axiomatización que nos dice que ciertas frases de ciertos discursos son o no buenos descriptores de esos discursos.

La representación del discurso contiene al discurso y algunas ideas acerca de su contenido. Aquí es donde interviene las reglas de estilo de Williams. Usaremos sus reglas para asociar tópicos al discurso.

Esto va a permitir tener criterios para determinar el éxito y el fracaso en la extracción de información relevante. Poder distinguir del conjunto de los

descriptores, aquellos que son candidatos a ser catalogados como buenos descriptores, o quizás una frase que resume el texto.

Algunas alternativas que pueden tomarse en cuenta en esta tarea pueden ser:

- Usar técnicas de modelos probabilísticos para detectar relevancia en el contexto.
- Uso de una red semántica, ontología del contexto.
- Permitir la interacción de un experto en el dominio del contenido del texto para comparar la relevancia obtenida por el sistema.
- Usar criterios los bibliográficos para asignar descriptores a documentos, por ejemplo CEPAL[17].

(C) Programar un parser para esta gramática

El parser o análisis de la estructura gramatical del lenguaje, debe implementar la gramática definida formalmente en el primer paso. Se debe usar un lenguaje declarativo, como Prolog, para evitar desviar la atención hacia el procesamiento de las reglas gramaticales.

(D) Escoger un dominio de conocimiento

El dominio debe ser escogido con cuidado, pues se tiene como hipótesis que el texto sigue las reglas de estilos expuesta anteriormente. Escoger cuidadosamente el contexto o dominio de aplicación quiere decir que:

- Se tiene un dominio cerrado que puede controlarse.
- Se tiene una disponibilidad de los documentos, están en formato digital y son de dominio público.
- Contamos con la ayuda de un experto en el dominio de conocimiento, del cual se extraerá su conocimiento, es decir la teoría-de-apoyo Además el experto puede ayudar a verificar los resultados obtenidos con la herramienta.

(E) Probar el sistema y evaluar los resultados

En la fase de prueba se deben diseñar varios experimentos para obtener resultados que puedan ser validados. Por supuesto, estos experimentos deben recibir como datos de entrada documentos del dominio anteriormente escogido.

Se tendrá por lo menos dos (2) grupos de control:

- Grupos de descriptores determinados por humanos.
- Grupos de descriptores determinados por humanos pero asistidos por la

[17] El concepto de descriptor temático de CEPAL es el siguiente: "Términos formados por una o más palabras claves que resumen o denotan un concepto, extraídos de un tesauro o vocabulario controlado utilizado por la unidad de información".

herramienta.

En [van Rijsbergen. C.J.] se explica un método convencional de evaluación en IR, comúnmente usado en evaluación. Para poner el problema de la evaluación en perspectiva se deben plantear tres preguntas: (1) ¿por qué evaluar? (2) ¿qué evaluar? (3) ¿cómo evaluar?.

La respuesta a la primera pregunta es principalmente social y económica. La parte social es bastante intangible, pero se relaciona principalmente con el deseo de poner una medida en las ventajas (o las desventajas) de tener un sistema de recuperación de información. Por ejemplo, qué ventaja los usuarios obtendrán (o qué daño será hecho) substituyendo las fuentes tradicionales de la información por un sistema de extracción completamente automática e interactivo?. Para algunas clases de sistemas de extracción la ventaja se puede medir más fácilmente que para otras. La respuesta económica asciende a la declaración de ¿cuánto va a costarle para utilizar uno de estos sistemas?, y además es ¿justo el costo?. Entonces si vale o no depende del usuario individual.

La segunda pregunta (¿qué evaluar?) considera los aspectos que determinan la capacidad del sistema de satisfacer al usuario. Desde 1966, Cleverdon [Cleverdon, C.W.] dio una respuesta a esto y enumeró seis cantidades mensurables principales. Sin embargo para efectos de este trabajo de investigación y en las actuales circunstancias tecnológicas, mas de 30 años después cuando se trata de medir la eficacia del sistema en función de la satisfacción del usuario, solo resultan relevantes las siguientes cantidades:

(1) la precisión del sistema, es decir, la proporción del material extraído que es realmente relevante.
(2) el tiempo de respuesta, es decir, el intervalo medio entre el tiempo que se hace la petición de la búsqueda y el tiempo en que se da una respuesta;
(3) la forma de presentación de la entrada y la salida que puede implicar un esfuerzo por parte del usuario en la obtención de respuestas;

La pregunta final (¿cómo evaluar?) tiene una respuesta técnicamente grande. Habría que examinar el concepto de relevancia, la cual es una noción subjetiva. Los humanos pueden diferenciar sobre la relevancia o la no-relevancia del contenido de los documentos. Esta noción de la relevancia ha sido explicada por Cooper [Cooper, W.S.] y la llama correctamente "relevancia lógica". La importancia esta definida en términos de la consecuencia lógica. Un documento es relevante a una necesidad de información si y solamente si contiene por lo menos una sentencia que sea relevante a esa necesidad. Lo que se quiere es tener una estrategia genérica para extraer los descriptores relevantes sin importar independientemente de una necesidad especifica de información.

Será importante entonces en este trabajo considerar metodológicas de evaluación

de sistemas NLP, pero estos métodos pueden resultar muy costosos y complejos. Margaret King destaca la importancia de los resultados de las evaluaciones a sistemas de NLP como datos invalorables, pero advierte que las evaluaciones varían enormemente en función del propósito, del alcance, y de la naturaleza de los objetos que están siendo evaluados [King, M.].

5. Conclusiones.

La lingüística computacional, no es solamente un método sino un paradigma con un esquema computacional del procesamiento del lenguaje. Este planteamiento ha dado lugar una amplísima diversidad de teorías lingüísticas, las cuales fueron expuestas en el marco teórico de esta propuesta.

El problema planteado en esta propuesta trata básicamente el procesamiento del lenguaje natural de documentos. Este procesamiento puede necesitar el conocimiento lingüístico del lenguaje y el conocimiento del mundo que maneja el hablante. El conocimiento lingüístico, en particular la morfología y sintaxis, puede ser expresados en forma de gramática para reconocer los constituyentes y la estructura de las oraciones. Esto se ha hecho durante años con diferentes teorías gramaticales. También se han realizado numerosos intentos por expresar la lingüística de un lenguaje con técnicas y modelos no simbólicos, tales como los modelos probabilísticos y los conexionistas. Los modelos conexionistas tienen limitaciones y solo han sido desarrollado en aplicaciones de reconocimiento del habla. Los modelos estadísticos por su parte realizan el estudio de una muestra (corpus) para modelar el uso de lenguaje en dicha muestra, pero al final se obtienen reglas gramaticales con probabilidades asociadas. Esto se hace con el fin de reducir la ambigüedad (al contar con las opciones mas probables), y mejorar la eficiencia de los sistema de NLP (los tiempos de respuesta mejoran al descartar opciones en función de su probabilidad). Entonces independientemente del modelo, simbólico o estadístico, el procesamiento del lenguaje natural involucra alguna gramática.

Las gramáticas simbólicas no pueden incluir toda la información necesaria para tratar el lenguaje natural. Las gramáticas probabilísticas tienen el atractivo de incorporar tanto el conocimiento como el uso, mientras que las gramáticas estándares solo manejan la competencia. Estas limitaciones han creado la tendencia de desarrollar sistemas híbridos que combinen ambos esquemas.

Por otra parte, la complejidad de los problemas tratados ha obligado a dividir y a especializar cada parte del sistema, y como consecuencia de ello han surgido técnicas y herramientas que faciliten el desarrollo de tales sistemas: lenguajes simbólicos declarativos, formalismos gramaticales y técnicas de parsing eficientes. Estas herramientas han marcado la diferencia en el desarrollo de los sistemas de procesamiento del lenguaje natural.

Como se explico en la definición del problema, lo que se propone es desarrollar una herramienta que permita asignar descriptores a un documento en idioma español. El grupo de descriptores pueden reemplazar la búsqueda exhaustiva sobre todo el texto y facilitar procesos propios de recuperación de información como por ejemplo la categorización y minería de texto, etc.

Como en cualquier ciencia aplicada, hay cierta separación entre lo esperado teóricamente y los resultados prácticos. El funcionamiento real de los sistemas NLP depende de muchas variables que generalmente contradicen las demostraciones teóricas. Entonces para determinar el éxito o no de esta propuesta, se debe considerar cuales son estas variables y su influencia, además del uso práctico e importancia teórica de la aplicación desarrollada.

Referencias

[Abney, S.] Abney, S., "Statistical Methods and Linguistics", en Klavans y Resnik (1996).
[ALC] "Association for Computational Linguistics". http://www.aclweb.org/
[Allen, J.] Allen, J. "Natural Language Understanding". Redwood City. Benjamin/Cummings. (1995).
[Bach, E.] Bach, E., "Syntactic Theory", Nueva York, Holt, Rinehart & Winston (1974).
[Barber, A.S.] Barber, A.S., Barraclough, E.D. and Gray, W.A. 'On-line information retrieval as a scientist's tool', Information Storage and Retrieval, 9, 429-44- (1973).
[Black, Garside y Leech] Black E., Garside G. y Leech G. "Statistically-driven Computer Grammmars of English". The IBM/Lancaster Approach. Rodopi B.V. 1993.
[Bod R. y Scha, R.] Bod R. y Scha, R., "Data-Oriented Language Processing: An Overview", ILLC Technical Report LP-96-13. Universidad de Amsterdam.
[Charniak, E.] Charniak, E., "Statistical Language Learning", Cambridge, The M.I.T. Press. (1993).
[Chomsky, N.] Chomsky, N. "Syntactic structures". 1957. La Haya, Muton.
[Church y Mercer] Church y Mercer, "Introduction to the Special Issue on Computacional Linguistics using large corpora", en Computational Linguistics, 19(1), pp. 1-24. (1993).
[Cleverdon, C.W.] Cleverdon, C.W. 'Progress in documentation. Evaluation of information retrieval systems', Journal of Documentation, 26, 55-67, (1970).
[Codd, E.] Codd, E., "How About Recently" (English Dialog with Relational Data Bases Using Rendezvous Version 1), en Shneiderman. 1978.
[Cooper, W.S.] Cooper, W.S., "A definition of relevance for information retrieval", Information Storage and Retrieval, 7, 19-37 (1971).
[Cormack, R.M.] Cormack, R.M., "A review of classification", Journal of the Royal Statistical Society, Series A, 134, 321-353 (1971).
[Covington, M.] Covington, Michael A. "Natural Language Processing for Prolog Programmers". Artificial Intelligence Programs The University of Georgia Athens, Georgia. PRENTICE HALL, Englewood Cliffs. NewJersey 07632.
[Cowie y Lehnert] Cowie, J. y Lehnert, W., "Information Extraction", Communications of the ACM. Enero 1996, Vol. 39, No. 1.
[Croft y Lewis] Croft, W. B. y Lewis, D. D., "An approach to natural language processing for document retrieval", in Proceedings ACM SIGIR International Conference on Research ans Development in Information Retrieval. New Orleans, LA (1987) 26-32.
[DiMarco y Hirst] DiMarco, C. y Hirst, G., "A computational theory of goal-directed style in syntax", Computational Linguistics. 19, 3 (Septiembre 1993), 451-499.
[Elmasri y Navathe] Elmasri R. y Navathe S. "Fundamentals of Database Systems", Seconds Edition, Addison-Wesley Publishing Company, Inc. Reading, Massachusetts, 1997.

[Fagan, J.] Fagan, J. "Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods", disertacion de doctorado, Departamento de ciencias de la computacion, Cornell University, 1987.
[Fagan, J.] Fagan, J., "Experiments in automatic phrase indexing for document retrieval: a comparison of syntactic an non-syntactic methods". Ph.D. Thesis. Tech. Report 87-868. Computer Science Department, Cornell University, Ithaca, NY (1987).
[Gazdar y Mellish] Gazdar, G. Mellish, C. "Natural Language Processing in PROLOG". 1989. Reading. Addison-Wesley.
[Grice, H.] Grice, H., P. "Logic and conversation". 1975. Cole, P. y J. Morgan, EDS., Syntax and semantics, vol. 3, 41-58, New York: Academic Press.
[Grishman, R.] Grishman, R. "Computational Linguistics: an introduction". Cambridge, Cambridge University Press. 1986
[Harris, L.] Harris, L. "The ROBOT System: Natural Language Processing Applied to Data Base Query", Processing of the ACM National Conference, 1978.
[Hendrix, G.] Hendrix, G., Sacerdoti, D., Sagalowicz, D. y Slocum, J. "Developing a Natural Language Interface to Complex Data", TODS, 1978.
[IBM-1] Intelligent Miner for Text "Text Analysis Tools version 2.2", Segunda Edicion, Junio 1998.
[IBM-2] Intelligent Miner for Text "TextMiner: Programming Interfaces version 2.2", SH12-6307-01. Segunda Edicion, Junio 1998.
[Jacobs y Rau] Jacobs, Paul y Rau, Lisa. "Innovations in text interpretation". Artificial Intelligence 63. (1993).
[Kay, M.] Kay, M., "Parsing in functional unification grammar", (1979;1985), publicado en Dowty, D., Kartunnen, L. y Zwicky, A. (1985) pp 251-278.
[King, M.] King, M., "Evaluating Natural Language Processing Systems", Communications of the ACM. Enero 1996, Vol. 39, No. 1.
[Klopp] von Klopp "A Practical Introduction to Prolog and Computational Linguistics".
[Knight, K.] Knight, Kevin. "Mining Online Text". Communications of the ACM. Noviembre 1999/Vol. 42 No. 11.
[Krovetz y Croft] Krovetz, R. y Croft, B. "Lexical Ambiguity and Information Retrieval", en TOIS, 1992.
[Kubácek, L.] Kubácek, L., "Confidence Limits for Proportions of Linguistics Entities", en Journal of Quantitative Linguistics, vol. 1, num. 1, pp. 56-61. (1994).
[Lancaster, F. W.] Lancaster, F.W., Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York (1968).
[Lewis y Sparck] Lewis, David D. y Sparck Jones, Karen., "Natural Language Processing for Information Retrieval", Communications of the ACM. Enero 1996, Vol. 39, No. 1.
[Locke y Booth] Locke W.N. y Booth A.D., "Machine Translation of Languages", Technology Press of MIT ans Wiley, Cambridge, Mass., 1955.
[Lyons, J.] Lyons, J., "Introduction to Theorical Linguistics", Cambridge, Cambridge University Press. (1968).

[Manaris y Slator] Manaris, Bill Z. y Slator, Brian M., "Interactive Natural Language Processing: Building on Success", Computer, IEEE, 1996
[McCARN, D.B.] McCARN, D.B. y LEITER, J., 'On-line services in medicine and beyond', Science, 181, 318-324 (1973).
[Minsky, M.] Minsky, M, Semantic Information Processing, MIT Press, Cambridge, Massachusetts (1968).
[Moreno, A.] Moreno Sandoval, Antonio. "Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos". Madrid. Editorial Sintesis. 1998.
[Ochoa, J.] Ochoa A., Jesús, I., Tutor: Davila, J.,"Formulacion de Normas Logicas para el discurso escrito de noticias", Tesis de Grado de Ingenieria de Sistemas, ULA - Venezuela. (Octubre 1999).
[Pollard y Sag] Pollard, C. y Sag, I., "An information-based approach to syntax and semantics", (1987), volume 1 Fundamentals. Chicago, Chicago University Press.
[Roche y Schabes] Roche, E. y Schabes, Y., "Finite-State Language Processing", Cambridge, The M.I.T. Press. (1997).
[Rumelhart y MacCleland] Rumelhart, D. y MacCleland y el grupo PDP. “Parallel Distributed Processing: Explorations in the Microstructure of Cognition”, Vol.1. Foundations. Vol 2. : Phychological an Biological Mpdels. Cambridge, M.I.T.
[Salton y Buckley] Salton, G. y Buckley, C. "Global Text Matching for Information Retrieval", en Science, 253, 1991.
[Salton, G.] Salton, G., Automatic Information Organization and Retrieval, McGraw-Hill, New York (1968).
[Salton, G.] Salton, G., 'Automatic text analysis', Science, 168, 335-343 (1970).
[Shannon y Weaver] Shannon, C.E. y Weaver, W, The Mathematical Theory of Communication, University of Illinois Press, Urbana (1964).
[Shieber, S.] Shieber, S. "An Introduction to unification-based approaches to grammar". 1986. Chicago, Chicago University Press. [Trad. esp.: Introducción a los formalismo gramaticales de unificación, Barcelona, Teide, 1989]
[Sparck, J.] Sparck Jones, K, "Some thoughts on classification for retrieval", Journal of Documentation, 26, 89-101 (1970).
[Toshinori, M] Toshinori, Munakata. "Knowledge Discovery". Communications of the ACM. Noviembre 1999/Vol. 42 No. 11.
[Uszkoreit, H.] Uszkoreit, H., “Mathematical Methods: Overview”, en The State of the Art of Human Language Technology, capítulo 11.1., 1996.
[van Rijsbergen. C.J.] van Rijsbergen, C. J., "Information Retrieval" Second Edition (London: Butterworths, 1979).
[Waldrop, M.] Waldrop, M. “Complexity: the emerging science at the edge of order and chaos”. New York, Touchtone. 1992.
[Wiebe, Hirst y Horton] Wiebe, J., Hirst, G. y Horton, D., "Language Use in Context", Communications of the ACM. Enero 1996, Vol. 39, No. 1.
[Williams, J.] Williams, Joseph. M. "Style Toward Clarity and Grace". The University of Chicago Press. Chicago and London.
[Wilson, B.] Wilson, Bill. "The NLP Dictionary". 1998 http://www.cse.unsw.edu.au/*billw/nlpdic.html

[Winograd, T.] Winograd, T. "Language as a Cognitive Process: Syntax". Reading, Addison-Wesley. 1983.
[Winograd, T.] Winograd, T., Understanding Natural Language, Edinburgh University Press, Edinburgh (1972).