

Tema 1

Introducción al Procesamiento del Lenguaje Natural

Ingeniería del Lenguaje Natural

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

<http://www.dlsi.ua.es/asignaturas/iln>



Índice

1. El lenguaje, las lenguas y su procesamiento automático.
2. Aproximación histórica.
3. Fundamentos filosóficos.
4. Módulos lingüísticos.
5. Arquitectura general de un sistema de PLN.
6. El problema de la ambigüedad.
7. Aplicaciones.
8. Fuentes de información actuales.

El lenguaje

- El lenguaje es uno de los aspectos fundamentales del comportamiento humano.
 - Principal vehículo de comunicación.
 - Socialización.
 - Razonamiento.
 - Escrito: transmisión el conocimiento de una generación a otra.
- Lenguaje y lenguas naturales

Diversidad lingüística

1. “La razón de la sinrazón que a mi razón se face...”
2. “Es un derecho de los ciudadanos el no presentar documentos exigidos por las normas aplicables al procedimiento de que se trate o que ya se encuentren en poder de la administración actuante”
3. “Los hoteles cierran el mejor agosto en ocupación desde el 2000”
4. “Introducir primeramente la parrilla por la bandeja 1 hasta la posición de la figura”
5. “La excitación de unidades corticales simples depende la orientación del estímulo en una parte específica del campo visual”
6. “ok tkiero bs”.



Diversidad lingüística

- Actualmente existen más de 7000 lenguas.
 - Muchas de ellas en peligro de extinción.
 - Cada una es reflejo de una sociedad inteligente y única en sí misma, con una historia propia.
 - Construcción gramatical y semántica específica.
 - Reflejo de su categorización cognitiva de la realidad y el mundo.



Diversidad lingüística

- Constante evolución:
 - Pequeñas variaciones llevan a grandes cambios.
 - "patético", "álgido", etc.
 - Introducción de novedades continuamente:
 - "Chatear".
- Constante influencia de unas sobre otras
 - "shopping center".



Sociedad de la información

- Cantidad ingente de información disponible que el ser humano debe gestionar.
 - Internet.
- Necesidad de procesar la información para pasar a la sociedad del conocimiento:
 - Capacidad de crear conocimiento



Sociedad de la información

- Cada día el ser humano genera grandes cantidades de información mediante las lenguas naturales.
 - 267 terabytes de datos en internet (2003)
 - Más de 35 terabytes de texto en html.
 - 440 terabytes en mensajes de correo electrónico.
 - 8 terabytes en libros al año
 - 37 terabytes en periódicos y revistas
 - 95 terabytes en documentos (literatura gris)
 - 16 terabytes en cine, 6 en música (CD), 22 en vídeo (DVD).
- 1 terabyte = 1 biblioteca universitaria

Fuente: *How Much Information? 2003* <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>



Complejidad lingüística

- La lengua objeto de interés para:
 - lingüística, traducción, crítica literaria,
 - filosofía, antropología, pedagogía,
 - psicología, derecho, hermenéutica, retórica,
 - ingeniería en informática, etc.
- Cada disciplina estudia las lenguas desde diferentes puntos de vista.
 - Estudio de las lenguas en toda su complejidad.
 - Interrelación: cada una necesita de las otras para alcanzar sus objetivos.



Ingeniería informática y las lenguas

- Ingeniería Informática:
 - Inteligencia Artificial
 - Ingeniería del Lenguaje Natural
 - **Lingüística Computacional o Procesamiento del Lenguaje Natural**



Ingeniería informática y las lenguas

- Ingeniería del Lenguaje Natural. Área de la ingeniería que engloba a la LC o PLN en el proceso de creación de sistemas.
 - Reconocimiento del habla.
 - Lingüística Computacional o Procesamiento del Lenguaje Natural.
 - Aplicaciones.



Ámbitos de aplicación del PLN

- El procesamiento automático de textos es necesario para gran cantidad de ámbitos:
 - Académicos (científico):
 - Lingüística de corpus, humanidades computacionales, ciencia de la computación.
 - Económicos y sociales (ingeniería):
 - Interacción hombre máquina.
 - Gestión automática de información en formato textual.
 - Internet.



Objetivos del PLN

- Encontrar mecanismos computacionales que permitan comprender y generar textos en una lengua natural.
- Automatizar la facultad lingüística del ser humano.
- Diseñar modelos matemáticos que permitan la automatización del procesamiento lingüístico mediante ordenadores.
- Formalización de las teorías y modelos lingüísticos y su implementación en sistemas automáticos.
- ...



APROXIMACIÓN HISTÓRICA. ORÍGENES DEL PLN.



PLN y la máquina inteligente

- El gran desafío de la informática es desarrollar la máquina inteligente.
- El principal rasgo de inteligencia es lingüístico:
 - Test de turing.



Lenguaje y procesamiento simbólico

- 1900.
- Hipótesis: las lenguas naturales pueden ser tratadas con mecanismos computacionales.
 - Frege, Russell, Wittgenstein, Carnap, etc.
 - Razonamiento matemático basado en la lógica aplicado a las lenguas.
 - Consideración de las lenguas como sistemas formales sensibles al procesamiento automático



Lenguaje y procesamiento simbólico

- Desarrollo de tres planteamientos (Bird et al. 2007):
 - Teoría lenguajes formales
 - Lógica simbólica
 - Principio de composicionalidad



Lenguaje y procesamiento simbólico

- Desarrollo de tres planteamientos:
 - Teoría lenguajes formales:
 - Lenguaje entendido como el conjunto de cadenas aceptadas por un autómata.
 - Lenguaje independiente del contexto y autómata descendente.
 - Base de la sintaxis computacional
 - Lógica simbólica
 - Principio de composicionalidad



Lenguaje y procesamiento simbólico

- Desarrollo de tres planteamientos:
 - Teoría lenguajes formales
 - Lógica simbólica:
 - Lógica de primer orden y proposicional
 - Método formal para representación semántica no ambigua, la inferencia y la interpretación.
 - $\text{ver}(j, m)$
 - Principio de composicionalidad



Lenguaje y procesamiento simbólico

- Desarrollo de tres planteamientos:
 - Teoría lenguajes formales
 - Lógica simbólica
 - Principio de composicionalidad:
 - Correspondencia sintaxis y semántica.
 - Introduce al recursión:
 - $p = \text{ver}(j, m)$
 - $\neg p$



Los primeros sistemas de PLN

AÑOS 50

- Sistema de TA inglés-ruso, basado en la equivalencia de palabras.
 - Traducción muy rudimentaria palabra a palabra:
 - "The *spirit* is willing but the *flesh* is weak"
 - El **espíritu** es fuerte pero la **carne** es débil
 - El **vodka** es bueno pero la **carne** está podrida



Los primeros sistemas de PLN

AÑOS 50

- El GAT (Georgetown Automatic Translator), y el CETA (Centre d'études pour la Traduction Automatique).
 - Se hace patente la naturaleza de los problemas a tratar y las limitaciones tanto teóricas como técnicas.



AÑOS 60

- El informe ALPAC, en 1964, supuso un freno pero no un impedimento para el desarrollo de diversos sistemas.
- El PLN consistió principalmente en métodos de análisis de palabras clave y *pattern matching*, dando lugar a sistemas como:
 - BASEBALL de Green (1963),
 - ELIZA de Weizenbaum (1966)
 - SIR de Raphael (1968),
 - STUDENT de Bobrow (1968)



AÑOS 70

- Primeras interfaces en lengua natural a base de datos:
 - Sistema LUNAR de Woods
- Aparecen diversos analizadores que usan gramáticas incontextuales (CFG):
 - Sistema SAD-SAM de Lindsay (Schank75)
 - Basado en las Gramática Generativa de N. Chomsky
- SHRDLU (Winograd 72) sistema para enviar órdenes a un robot.
 - Basado en las relaciones funcionales entre palabras de Halliday: *Systemic Grammar*.



AÑOS 70

- Desarrollo por Woods de las Redes de Transición Aumentadas (ATN):
 - Mejora la potencia de las expresiones regulares y de las gramáticas incontextuales al incorporar restricciones.
 - Ej. Concordancia.
 - Permite que una ATN incorpore más información contextual cuando se genera un análisis.
 - Potencia la metodología de diseño "ad-hoc", donde cada nueva aplicación requiere una nueva ATN.



AÑOS 80

- Aparecen diversos formalismos que, además de contar con la potencia de las ATN, se basaban en estructuras teóricas más formales.
- En 1983, Chomsky propuso su Teoría de Rección y Ligamiento (*Government and Binding*).
 - Se da mayor importancia al léxico, que contiene toda la información léxico, semántica y sintáctica para la formación/análisis de la oración.
 - Se reduce el papel de la gramática a una serie de reglas de buena formación.



AÑOS 80

- En esta línea surgen una serie de gramáticas como las
 - Gramáticas de Estructura de Frase Generalizadas: GPSG (Sells 89)
 - Gramáticas Léxico Funcionales de Bresnan: LFG (Bresnan 82),
 - Gramáticas de Unificación Funcionales de Kay: FUG (Dowty 85).



AÑOS 80

- A partir de los trabajos de Colmerauer aparecen las gramáticas lógicas:
 - Gramáticas de Cláusulas Definidas de Pereira y Warren DCG (Pereira80).
- Aplicaciones:
 - Ariane-78, EUROTRA o ATLAS, en el campo de la Traducción Automática, y
 - TEAM (Grosz 87), CHAT-80 (Warren 82), ORBI (Pereira 82) en el campo de las interfaces con Bases de Datos



AÑOS 90

- Extensiones a formalismos ya introducidos en los años 80
 - Representación de las dependencias a larga distancia y las estrategias requeridas para el análisis y eliminación de la ambigüedad del texto.
 - Anáforas



AÑOS 90

- Cambio de interés de los principales organismos de I+D:
 - Años 60 se centraba en el control de procesos y las técnicas de programación,
 - Actualmente se centra en la Inteligencia Artificial y sus **aplicaciones**.



SIGLO XXI

- Paulatino abandono de sistemas de simbólicos (reglas manuales) y desarrollo de técnicas estadísticas.
 - Jelinek en IBM
- Sistemas a gran escala frente a los sistemas de pequeña escala.
- Desarrollo de métodos estadísticos, recuperación de información, uso de corpus de textos grandes y de diccionarios ya existentes.
 - Base para producir nuevos sistemas a gran escala con cierta rapidez.
 - Tendencia hacia el trabajo empírico.



FUNDAMENTOS FILOSÓFICOS



Dos posturas

- Debate filosófico (aún en vigor):
- Inicio S. SVII – XVIII (Ilustración).
 - El conocimiento NO proviene por revelación divina.
 - Racionalistas vs Empiristas.



Racionalistas

- Descartes, Leibniz, etc.
- El conocimiento tiene su origen en el razonamiento humano, en el pensamiento.
- Innatismo.
- Chomsky.



Empiristas

- Locke, etc.
- La fuente primaria de conocimiento es la experiencia a través de nuestras facultades sensitivas.
 - El razonamiento juega un papel, pero es secundario.
 - Ej: Heliocentrismo de Galileo, basado en la observación cuidadosa de los planetas.



El debate en Lingüística y PLN

- Las lenguas, ¿sin innatas o se basan en la experiencia?



El debate en Lingüística y PLN

- El Modelo racionalista: N. Chomsky.
 - Una parte considerable del conocimiento que se debe utilizar para el TL puede ser fijado de antemano.
 - Prescrito, codificado e incorporado como conocimiento inicial para cualquier proceso de TL.



El debate en Lingüística y PLN

- El Modelo empirista:
 - El conocimiento lingüístico se puede inferir a partir de la experiencia, que se puede recoger a través de corpus textuales.
 - Utilización de mecanismos como:
 - La asociación o la generalización: conocer una palabra por la compañía que lleva.
 - La distribución
 - Técnicas estadísticas y aprendizaje automático.



El debate en Lingüística y PLN

- Actualmente se buscan aproximaciones híbridas.
 - El ser humano nace con la capacidad innata de razonamiento analógico y métodos de aprendizaje que utiliza para identificar patrones semánticos por su experiencia sensible.
 - En PLN:
 - Desarrollo de reglas manuales (modelo racionalista)
 - Completado con información de corpus (modelo empirista)



Situación actual del PLN

- Almacenamiento masivo de información
- Técnicas de aprendizaje automático
- Anotación de corpus
- La evaluación



Situación actual

- Almacenamiento masivo de información
 - Uso de potentes algoritmo de indexación y búsqueda.
 - Poco uso de conocimiento lingüístico.
 - Ejemplo: Google.
- Técnicas de aprendizaje automático
- Anotación de corpus
- La evaluación



Situación actual

- Almacenamiento masivo de información
- Técnicas de aprendizaje automático
 - Basado en el modelo empirista.
 - Desarrollo de sistemas “reales”, pero
 - Imposibilidad de desarrollar sistemas con precisión del 100%
 - Parcialidad de las técnicas de aprendizaje (clasificadores, etc.).
 - Selección de rasgos de aprendizaje.
 - Error humano en las muestras de entrenamiento.
 - Creatividad: el carácter creativo del lenguaje hace imposible dar cuenta de todos los casos en el uso de las lenguas.
- Anotación de corpus
- La evaluación



Situación actual

- Almacenamiento masivo de información
- Técnicas de aprendizaje automático
- Anotación de corpus
 - Detallar la información lingüística en los corpus (sintáctica, semántica, etc.).
 - Manual, desarrollada por expertos.
 - Muestras de entrenamiento de los sistemas basado en aprendizaje automático
 - *Gold Standard*: Muestras de análisis correcto para la evaluación de sistemas
- Problema de la evaluación



Situación actual

- Almacenamiento masivo de información
- Técnicas de aprendizaje automático
- Anotación de corpus
- Problema de la evaluación
 - Necesidad de comparar los sistemas y evaluarlos de manera empírica
 - Necesidad de métricas objetivas y recursos.
 - Aparición de competiciones: TREC, CLEF, SENSEVAL, etc.



MÓDULOS LINGÜÍSTICOS. NIVELES DE DESCRIPCIÓN LINGÜÍSTICA.



Módulos lingüísticos

- **Módulo fonético y fonológico:**
 - Sonidos de la lengua
 - Fonemas: "casa" vs "pasa"
- **Módulo morfológico (morfoléxico):**
 - Unidades mínimas de las palabras: los monemas, morfemas y lexemas.
 - Flexión gramatical: "casa - casas"
"Deducir - deduje" ...
 - Composición y derivación.



Módulos lingüísticos

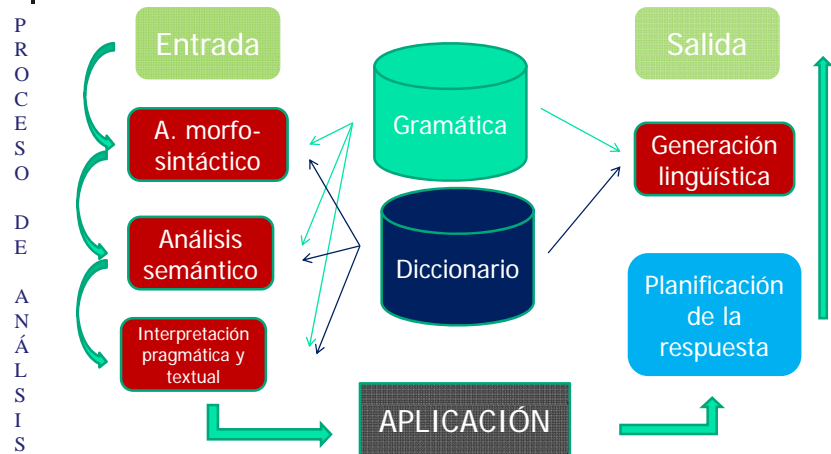
- **Módulo sintáctico:**
 - Combinaciones de palabras en estructuras superiores: sintagmas y oraciones.
- **Módulo semántico:**
 - Significado de las palabras: léxico-semántico
 - Significado completo de oraciones y textos.



Módulos lingüísticos

- **Nivel Textual:** cohesión y coherencia de los textos.
 - Anáfora, desarrollo temático, intencionalidad, etc.
 - No es un módulo en sí mismo.
- **Nivel Pragmático:**
 - Relación del texto con el contexto comunicativo: productor, receptor, mundo referencial, etc.
 - Aspectos que afectan directamente a la interpretación de las oraciones.
 - No es un módulo en sí mismo.

Sistema de PLN



El proceso de análisis

- **Análisis morfo-léxico o categorial:**
 - Tokenización: detección de palabras.
 - Stemmer: detección de la raíz de las palabras y rasgos morfológicos.
 - Clasificación de palabras por categoría gramatical.
 - Uso del diccionario y reglas morfológicas.
- **Análisis sintáctico:**
 - Especificación de las relaciones sintácticas entre las unidades léxicas:
 - Constituyentes
 - Dependencias.

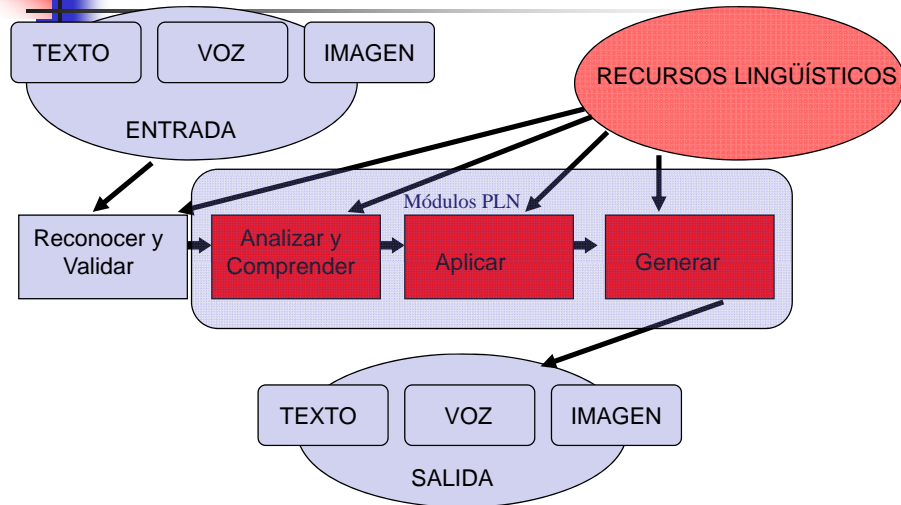
El proceso de análisis

- **Análisis semántico:**
 - **Semántico-léxico:**
 - Especificación del significado de las palabras.
 - **Semántico-oracional:**
 - A partir de la estructura sintáctica, generación de la forma lógica asociada que representa el significado o sentido de la oración.

El proceso de análisis

- **Análisis textual:**
 - Combinación las interpretaciones de las oraciones para determinar el tema e intención del texto, etc. (según la aplicación).
 - Análisis de fenómenos textuales: anáfora, marcadores del discurso, tema-remata, etc.
- **Análisis contextual o pragmático:**
 - Interpretación final del texto, en función de las circunstancias del contexto comunicativo.

ARQUITECTURA GENERAL DE SISTEMA ILN



ARQUITECTURA GENERAL DE SISTEMA ILN

- Reconocimiento del habla
- Análisis, comprensión y generación de la lengua
- Aplicación

ARQUITECTURA GENERAL DE SISTEMA ILN

- Reconocimiento del habla
 - Reconocimiento y síntesis de voz
 - Objetivo: Traducir la entrada hablada en una salida escrita-digital → separación de palabras, reconocimiento de fonemas, etc.
- Análisis, comprensión y generación de la lengua
- Aplicación

ARQUITECTURA GENERAL DE SISTEMA ILN

- Reconocimiento del habla
- Análisis, comprensión y generación de la lengua
 - Procesamiento del Lenguaje Natural (PLN)
 - Análisis léxico, morfológico, sintáctico, semántico y contextual de la lengua
 - Comprensión conceptual del lenguaje.
 - Generación del lenguaje
- Aplicación



ARQUITECTURA GENERAL DE SISTEMA ILN

- Reconocimiento del habla
- Análisis, comprensión y generación de la lengua
- Aplicación
 - Sistemas de Extracción de información.
 - Sistemas de recuperación de información.
 - Sistemas de búsqueda de respuestas.
 - Sistemas de diálogo e interacción hombre-máquina.
 - Traducción automática.
 - Resúmenes automáticos...



LA AMBIGÜEDAD LINGÜÍSTICA EN PLN



El Problema de la Ambigüedad

- El principal problema del tratamiento del lenguaje es la **ambigüedad**.
- Toda palabra u oración es ambigua si permite más de una interpretación.
- Tipos:
 - Ambigüedad léxica (categorial o semántica)
 - Ambigüedad sintáctica
 - Ambigüedad semántica
 - Ambigüedad textual
 - Ambigüedad pragmática: referencial



El Problema de la Ambigüedad

- El tratamiento de la ambigüedad contempla dos sub-problemas:
 - La **representación** del problema: cómo las diversas interpretaciones se representan en un sistema.
 - La **interpretación** del problema: qué estrategias se siguen cuando aparece una ambigüedad para determinar una u otra interpretación.



Ambigüedad léxica

- Una palabra tiene más de un significado.
- Tipos:
 - Ambigüedad léxico-semántica:
 - La que afecta sólo al nivel semántico.
 - Ejemplos:
 - Juan dejó el periódico en el **banco**.
 - Se sentó en el **banco**.
 - Entró en el **banco** y fue a la ventanilla.
 - El avión localizó el **banco** y comunicó su situación.



Ambigüedad léxica

- Una palabra tiene más de un significado.
- Tipos:
 - Ambigüedad léxica categorial:
 - La palabra pueden pertenecer a más de una categoría gramatical (nombre, verbo, etc.).
 - Afecta morfo-sintáctico y semántico.
 - Ejemplos:
 - "El cura impartió el santo sacrificio."
 - "La cura será muy dolorosa."
 - "El médico cura al enfermo."



Resolución ambigüedad léxica-categorial

- *Part of speech tagger* (PoS tagger): determinan la categoría gramatical a partir del contexto.
- Principales categorías: nombre, verbo, adjetivo, determinante, preposición y adverbio.

*Este enfermo no tiene **cura#N***

*Este médico **cura#V** sin dolor*


*El **cura#N** dirá misa a las 12:00*



Resolución ambigüedad léxica-semántica

- Asignar el sentido correcto a las palabras:
 1. Resolución de la ambigüedad categorial
 2. Determinar el sentido correcto, pero ¿qué es un sentido?

«Te voy a firmar la cara con la **planta(?)** de mi pié.»



Resolución ambigüedad léxica-semántica

- Asignar el sentido correcto a las palabras

«Te voy a firmar la cara con la **planta(?)** de mi pié.»

WordNet 1.5

1. **planta, piso** -- a room or set of rooms comprising a single level of a multi-level building
2. **planta, flora** -- a living organism lacking the power of locomotion
3. **planta** -- the underside of the foot
4. **planta, fábrica** -- buildings for carrying on industrial labor
5. **planta, distribución** -- a floor plan for the ground level of a building



Ambigüedad sintáctica

- Ambigüedad sintáctica
 - La vendedora de periódicos del barrio.
 - Juan vio al ladrón con los prismáticos
 - Pedro vio a Juan en lo alto de la montaña con los prismáticos



Ambigüedad sintáctica

- Ambigüedad **sintáctica-estructural**:
 - La oración tiene diferentes estructuras sintácticas.
 - Se genera más de un árbol sintáctico de derivación.
- Tipos:
 - De origen preposicional (*pp-attachment*).
 - De origen coordinativo.
 - Por composición de nombres.

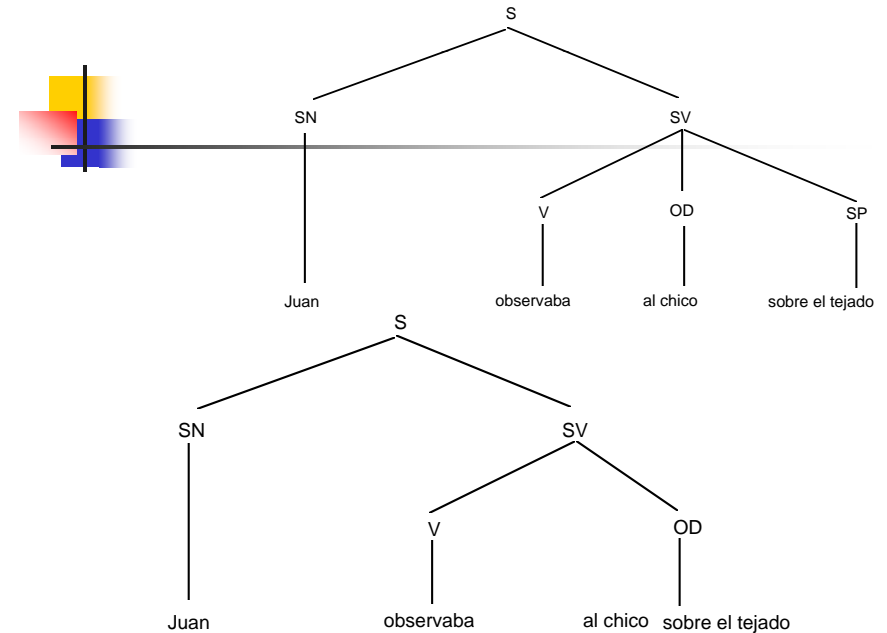


Ambigüedad Estructural

- Ambigüedad estructural de origen coordinativo:
 - *Juan o Víctor y Elena deberían ir*

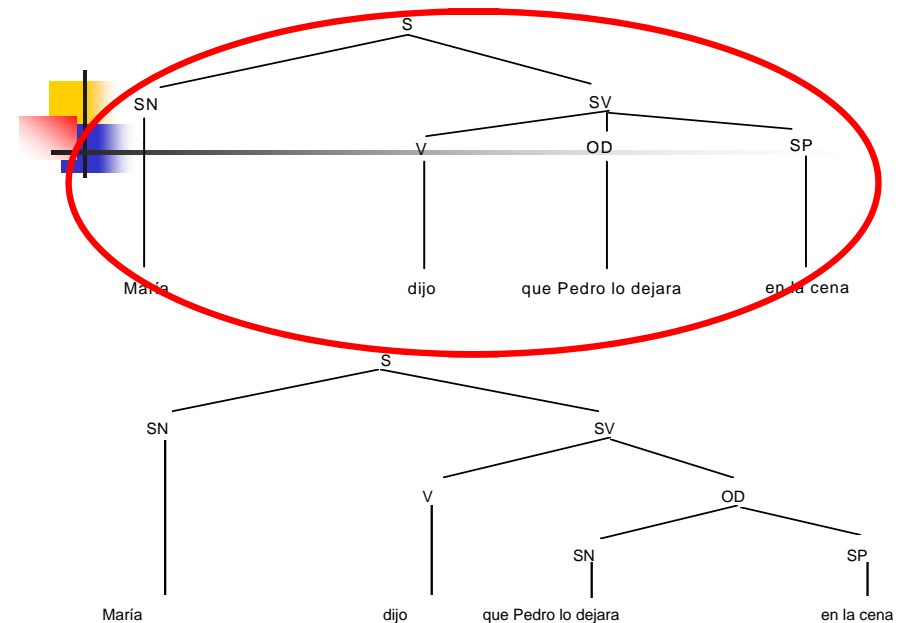
Ambigüedad Estructural

- Ambigüedad estructural de origen preposicional (*pp-attachment*):
 - Múltiples ligamientos que puede tener un sintagma preposicional en una oración.



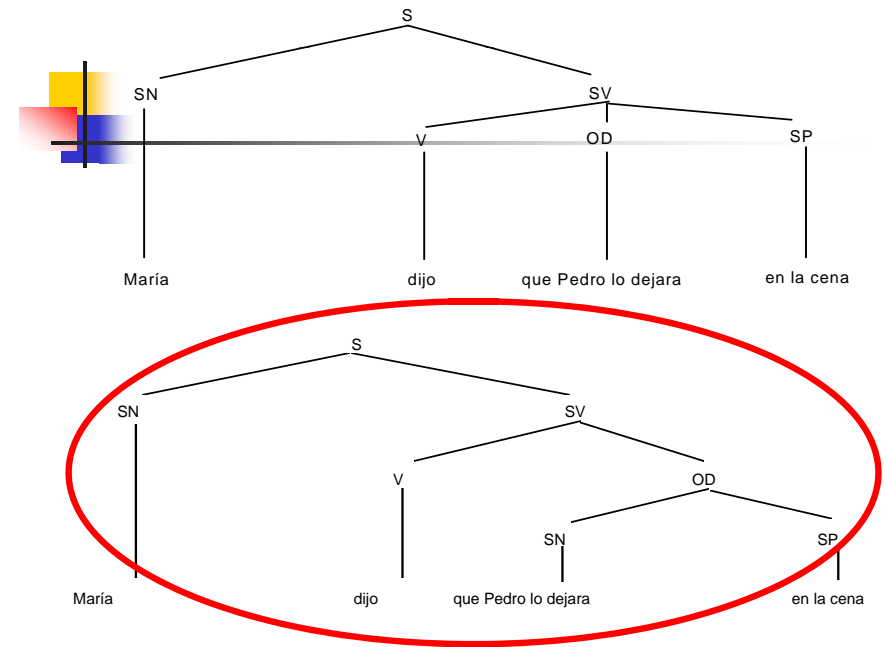
Ambigüedad Estructural

- Ambigüedad estructural de origen preposicional.
 - Un SP puede unirse a más de un nodo
 - Soluciones I:
 - El principio del mínimo ligamiento:
 - Consiste en la preferencia por el análisis sintáctico que genere un número menor de nodos en el árbol de análisis.



Ambigüedad Estructural

- Ambigüedad estructural de origen preposicional.
 - Soluciones II:
 - El principio de la asociación a derechas o última clausura:
 - Consiste en las preferencias de los nuevos constituyentes oracionales por ser interpretados como parte del actual elemento oracional en construcción, en vez de parte de algún constituyente superior en el árbol de análisis.

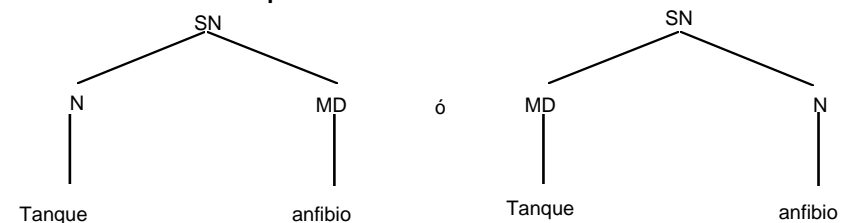


Ambigüedad Estructural

- Ambigüedad estructural por composición de nombres:
 - Este tipo de ambigüedades aparecen por la **unión de sustantivos**, debido a que en un nivel puramente sintáctico cualquiera de los sustantivos puede funcionar de núcleo actuando los demás de modificadores.

Ambigüedad Estructural

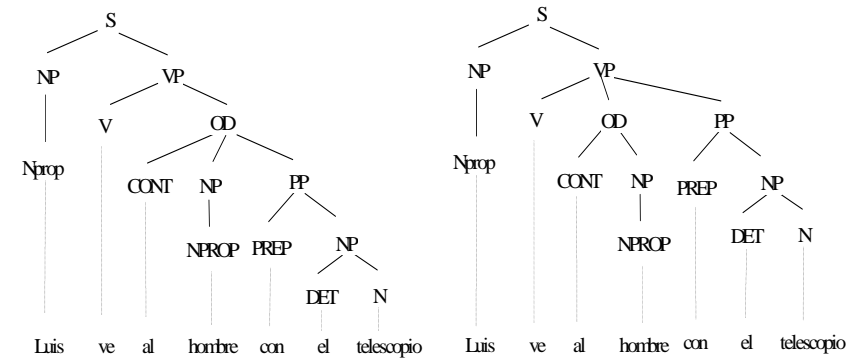
- Ambigüedad estructural por composición de nombres:
 - "Hombre Rana", "Hombre jardín" o "Tanque anfibio", ...



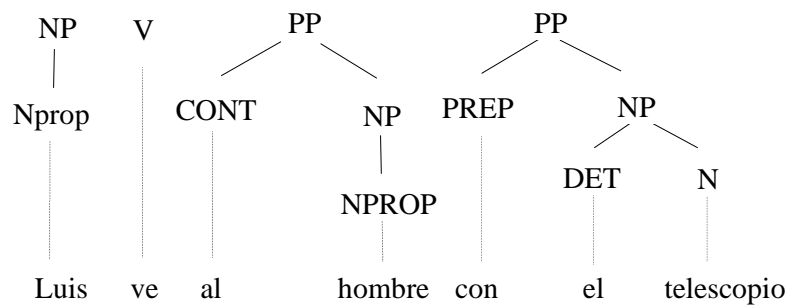
Análisis sintáctico

- Dos opciones:
 - Análisis sintáctico total: profundo.
 - Análisis sintáctico parcial (*chunker*): superficial.

Análisis sintáctico total



Análisis sintáctico parcial



Ambigüedad Semántica

- Una oración puede tener más de una forma lógica asociada.
- Ejemplo: Ambigüedad en el ámbito de la cuantificación.
 - Aparecen en una misma oración un cuantificador existencial y un cuantificador universal:
 - "El profesor recibió un regalo de todos los padres"
 - "Un juez decidió en cada sentencia"



Ambigüedad Semántica

- *"El profesor recibió **un** regalo de **todos** los padres"*
 - ¿Un único regalo de todos los padres?
 - El cuantificador existencial toma mayor ámbito que el universal.
 - ¿Un regalo de cada padre?
 - El cuantificador universal toma mayor ámbito que el existencial.



Ambigüedad de ámbito de cuantificación

- En [Dahlgreen89] se indica:
 - Existe una tendencia inherente a que los cuantificadores universales tomen un mayor ámbito que los existenciales.
 - Excepción:
 - "Pedro prefiere una mujer con todos los encantos"



Ambigüedad textual

- Diversas interpretaciones que puede tener una oración dependiendo del contexto.
 - Ejemplo:
 - "Luis Miranda tuvo que retocar el retrato de Felipe"
 - Para solucionarlas hay que conocer todo el texto donde aparece, no sólo la oración.



Ambigüedad textual

- Ejemplo. Solución en el contexto:
 - "Luis Miranda parece un buen artista. Felipe mi primo le encargo un retrato de su hija hace dos semanas. Fue a su taller a ver como había quedado y no estaba muy conforme. Luis Miranda tuvo que retocar el retrato de Felipe. Quizá no tenía que haber ido a un pintor tan modernista."



Ambigüedad textual

- Muchas ambigüedades de este tipo vienen originadas por una ambigüedad léxica.
 - "Me he dejado el periódico en el banco."
 - "Tengo el gato en casa."
- Aunque algunas ambigüedades léxicas pueden resolverse sólo con el contexto oracional:
 - "He ingresado el dinero en el banco esta mañana."
 - "Se me averió el gato y no cambié la rueda."



Ambigüedad textual y pragmática

- **Ambigüedad anafórica:**
 - Un pronombre o sintagma nominal definido pueden tener más de una antecedente.
 - "El hermano de Luís es un mecánico. **Él** arregló **su coche** en un día"
 - "Juan Bravo fue recibido por el director de SOL UNION S.A. **El fantástico hombre de negocios** por fin logró un éxito".



Ambigüedad pragmática

- **La anáfora:**
 - Mecanismo que permite hacer en un discurso una referencia abreviada a alguna entidad o entidades con la confianza de que el receptor del discurso sea capaz de interpretar la referencia y por consiguiente determinar la entidad a la que se alude (Hirst, 1981)
 - *Pedro, y María, son novios, pero él, no quiere casarse.*



El problema de la anáfora

RESOLUCIÓN Y GENERACIÓN

- Debe ser resuelta automáticamente por el sistema para poder interpretar correctamente el texto.



El problema de la anáfora

ESTRATEGIAS DE RESOLUCIÓN DE LA ANÁFORA

- Estrategias basadas en conocimiento lingüístico
- Estrategias basadas en corpus



El problema de la anáfora

ESTRATEGIAS DE RESOLUCIÓN DE LA ANÁFORA

- Estrategias basadas en conocimiento lingüístico
 - Imitan fuentes de conocimiento humano
 - Consultivos
 - una única fuente de información
 - Democráticos
 - combinan varias fuentes de información
 - mecanismos de restricciones y preferencias
 - reglas para descartar candidatos
 - reglas para ordenar los candidatos
- Estrategias basadas en corpus



El problema de la anáfora

ESTRATEGIAS DE RESOLUCIÓN DE LA ANÁFORA

- Estrategias basadas en conocimiento lingüístico
- Estrategias basadas en corpus



El problema de la anáfora

ESTRATEGIAS DE RESOLUCIÓN DE LA ANÁFORA

- Estrategias basadas en conocimiento lingüístico
- Estrategias basadas en corpus
 - Estudian corpus a través de herramientas estadísticas
 - Proponen modelos probabilísticos: Aprendizaje Automático.



El problema de la anáfora

ESTRATEGIAS DE RESOLUCIÓN DE LA ANÁFORA

- Estrategias basadas en conocimiento lingüístico
- Estrategias basadas en corpus



Ambigüedad y fases de análisis

Fase de análisis	Principales tipos de ambigüedad en PLN
Análisis morfo-sintáctico	Léxica-categorial, estructural.
Análisis semántico	Léxico-semántica, diversidad de formas lógicas.
Análisis textual - pragmático	Ambigüedad anafórica.



APLICACIONES



Ingeniería Lingüística

APLICACIONES

- Aplicaciones basadas en tratamiento textual
- Aplicaciones basadas en diálogos hombre-máquina

Ingeniería Lingüística

APLICACIONES

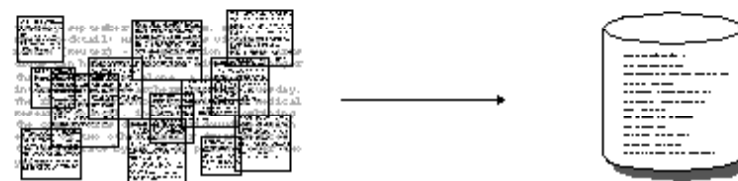
■ Aplicaciones basadas en tratamiento textual

- Extracción de Información (IE)
 - obtienen información relevante desde textos
- Recuperación de Información (IR)
 - seleccionan textos según algún requisito de consulta
- Búsqueda de Respuestas (QA)
- Traducción automática bilingüe/multilingüe
- Producción automática de resúmenes
- Corrección automática de textos
 - procesadores de textos
- Producción automática de textos

Ingeniería Lingüística

APLICACIONES

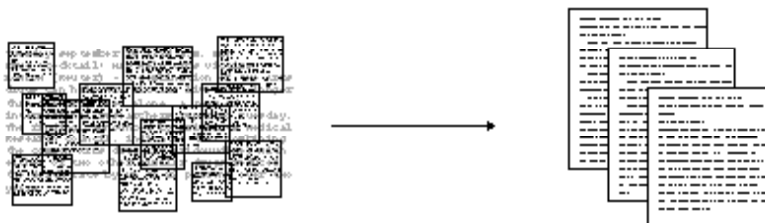
Extracción de información



Ingeniería Lingüística

APLICACIONES

Recuperación de información



Ingeniería Lingüística

APLICACIONES

- Aplicaciones basadas en tratamiento textual
- Aplicaciones basadas en diálogos hombre-máquina



Ingeniería Lingüística

APLICACIONES

- Aplicaciones basadas en tratamiento textual
- Aplicaciones basadas en diálogos hombre-máquina
 - Sistemas de diálogo
 - Automatización del comportamiento humano del diálogo
 - Formalización de aspectos intelectuales como:
 - Intenciones y deseos del usuario (emisión)
 - Conocimiento y creencias sobre el mundo (recepción)
 - Relación conocimiento-acción (acción)
 - Aplicaciones:
 - Orientados a tareas.
 - Orientados a la extracción y/o recuperación de información.



Ingeniería Lingüística

APLICACIONES

- Aplicaciones basadas en tratamiento textual
- Aplicaciones basadas en diálogos hombre-máquina



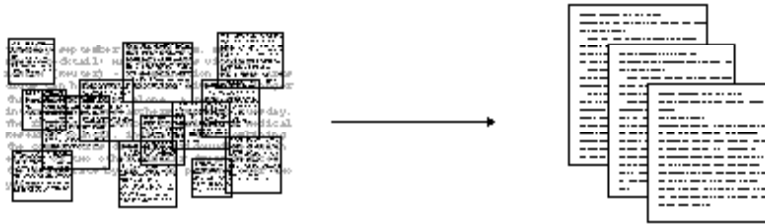
Sistemas de recuperación de información



Recuperación de información

- Documentos relevantes a una petición de información
- Tareas de un sistema de IR:
 - Indexación:
 - Representación eficiente de documentos
 - Tratamiento de preguntas:
 - Representación interna de la preguntas
 - Comparación de preguntas y documentos:
 - Medida de similitud entre preguntas y documentos

Recuperación de Información

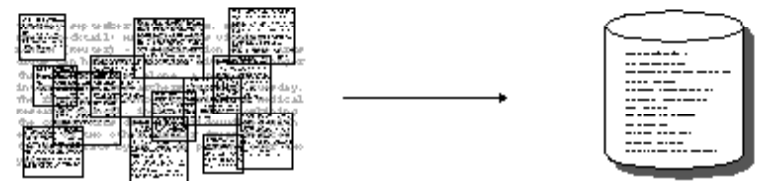


Sistemas de extracción de información

Extracción de Información

- Cowie y Lehnert (1996). "Técnica que proporciona determinada información denominada relevante de un conjunto de textos todos ellos relevantes"
- Gaizauskas y Wilks (1998). "Es la actividad de extraer automáticamente un tipo de información pre-especificada desde textos"

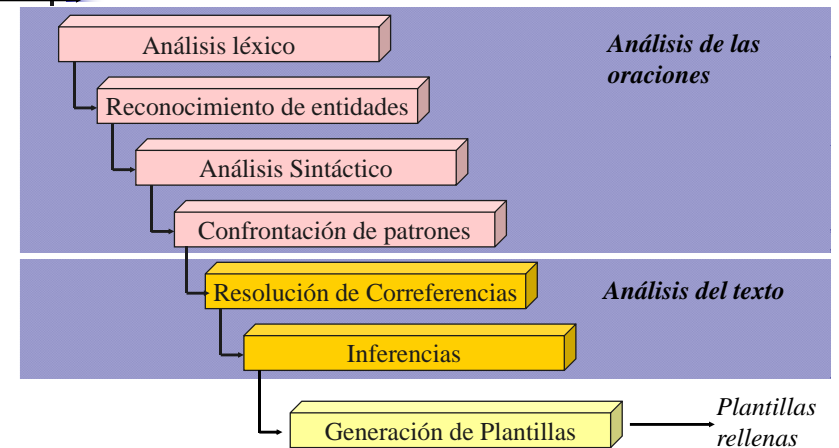
Extracción de Información



Extracción de Información

- Los sistemas deben encontrar y relacionar información relevante, e ignorar información NO relevante.
- La relevancia se determina a partir de guías predefinidas de dominio, las cuales deben especificar con la mayor exactitud posible el tipo de información a extraer.
- Desde la perspectiva del PLN, los sistemas de EI deben trabajar a distintos niveles: desde el reconocimiento de palabras hasta el análisis de oraciones y desde el entendimiento a nivel de oracional hasta el texto completo.

Extracción de Información



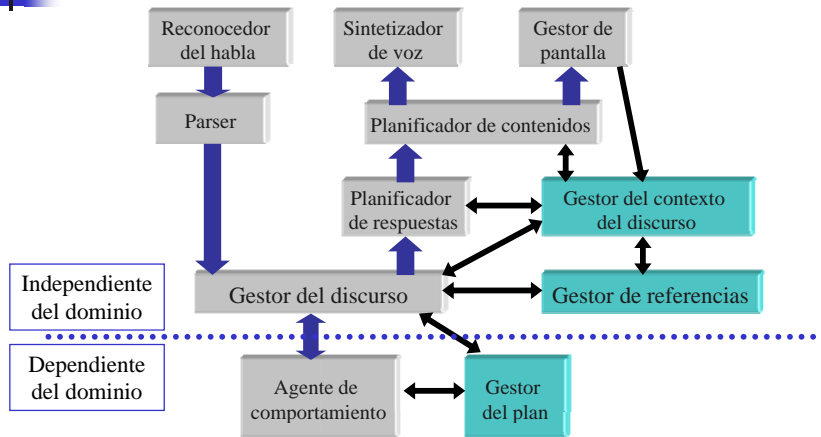
Sistemas de diálogo

Los sistemas de diálogo

- **Arquitectura básica** (Bernsen et al. 1998)
 - PLN (comprensión y generación)
 - Gestión del diálogo (control y contexto)
 - Procesamiento del habla (reconocimiento y síntesis)
- **PLN en los sistemas de diálogo** (Moreno et al. 1999)
 - análisis léxico
 - análisis morfológico
 - análisis sintáctico
 - análisis semántico
 - análisis contextual

Los sistemas de diálogo

ARQUITECTURA (Allen et al. 2001)



Fuentes y organizaciones

Fuentes y organizaciones

CONGRESOS ESPECÍFICOS

- ACL: *Association for Computational Linguistics*.
 - EACL: European chapter of the *Association for Computational Linguistics*
 - NAACL: North American chapter of the *Association for Computational Linguistics*
- COLING: *International Conference on Computational Linguistics*
- RANLP: *Recent Advances in Natural Language Processing*
- SEPLN: *Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*

Fuentes y organizaciones

REVISTAS

- *Computational Linguistics*.
- *Journal of Artificial Intelligence Research*
- *Artificial Intelligence*
- *Computing and Humanities*
- *ACM of Communications*
- *Machine Translation*
- *Procesamiento del Lenguaje Natural*
- *Revista Iberoamericana de Inteligencia Artificial*



Fuentes y organizaciones

DIRECCIONES DE INTERÉS

- SEPLN- www.sepln.org
- ACL- www.aclweb.org
- COLING - www.coling.org
- ...