



diplomatura universitaria en
inteligencia artificial



FACULTAD DE CIENCIAS
EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES

Procesamiento de Lenguaje Natural

Análisis Sintáctico

Qué vamos a ver hoy?

Análisis Sintáctico

- Qué es el Análisis Sintáctico?
- Categorización y etiquetado de palabras.
 - Etiquetado POS.
- Tipos de parsing.
 - Superficial.
 - De dependencia.
 - Constituency.



Londres es la capital y
mayor ciudad de Inglaterra y
del Reino Unido.



londres unido ciudad
 reino mayor
 capital
 inglaterra



Londres es la capital y
mayor ciudad de Inglaterra y
del Reino Unido.



londres unido ciudad
reino mayor
capital
inglaterra

**Una colección de palabras sin
ningún orden ni estructura!**



- La sintaxis y la estructura de las oraciones generalmente están relacionadas, y existe un conjunto de reglas, convenciones y principios específicos que rigen la forma en que se combinan las palabras.
- El conocimiento sobre la estructura y la sintaxis del lenguaje es útil en diversas áreas:
 - Procesamiento de texto, la anotación y el análisis para otras operaciones, como la clasificación o el resumen de texto.
- Se realiza luego del pre-procesamiento de texto.

Part-of-speech

Shallow parsing
chunking

Dependency
parsing

Constituency
parsing



Análisis Sintáctico

Qué tiene en cuenta?

A diferencia del análisis léxico tiene en cuenta:

- **Orden de las palabras y significado.**
 - Intenta determinar cómo son las dependencias entre las palabras.
 - Si se altera el orden de las palabras, puede dificultarse la comprensión de la oración.
- **Qué pasa con las stopwords?**
 - Si bien son comunmente eliminadas, para análisis sintáctico son implirantes.
 - Eliminarlas puede alterar el significado de una oración.
 - Por ejemplo: “no”.
- **Los POS.**
 - Identificar adecuadamente los POS es importante.
 - “cuts and bruises” (noun)
 - “an apple is cut” (verb)



Análisis Sintáctico

- La sintaxis y la estructura de las oraciones generalmente están relacionadas, y existe un conjunto de reglas, convenciones y principios específicos que rigen la forma en que se combinan las palabras.
- El conocimiento sobre la estructura y la sintaxis del lenguaje es útil en diversas áreas:
 - Procesamiento de texto, la anotación y el análisis para otras operaciones, como la clasificación o el resumen de texto.
- Se realiza luego del pre-procesamiento de texto.



Análisis Sintáctico

- La sintaxis y la estructura de las oraciones generalmente están relacionadas, y existe un conjunto de reglas, convenciones y principios específicos que rigen la forma en que se combinan las palabras.
- El conocimiento sobre la estructura y la sintaxis del lenguaje es útil en diversas áreas:
 - Procesamiento de texto, la anotación y el análisis para otras operaciones, como la clasificación o el resumen de texto.
- Se realiza luego del pre-procesamiento de texto.



Se encuentra en la intersección entre Análisis Léxico y Sintáctico. Puede ser considerado parte de cualquiera de las etapas.

dependen de gramáticas







Las palabras son la unidad más pequeña de un idioma.

Son independientes y tienen un significado propio.

Words

Las palabras son la unidad más pequeña de un idioma.

Son independientes y tienen un significado propio.



Phrases

Las palabras se pueden ordenar de manera que les den sentido, de manera que cada palabra pertenece a una categoría de frase correspondiente y una de las palabras es la principal de la oración.



Clauses



Words

Las palabras son la unidad más pequeña de un idioma.

Son independientes y tienen un significado propio.



Phrases

Las palabras se pueden ordenar de manera que les den sentido, de manera que cada palabra pertenece a una categoría de frase correspondiente y una de las palabras es la principal de la oración.



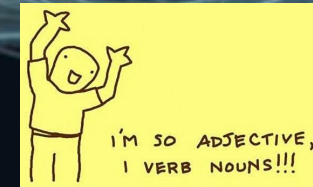
Clauses

Las cláusulas son grupos de palabras con alguna relación entre ellas, generalmente conteniendo un sujeto y predicado.

El sujeto puede ser una frase verbal o verbo con objeto.

Las cláusulas pueden actuar como oraciones independientes o varias cláusulas se pueden combinar para formar una oración.





- Las part-of-speech (partes del habla o discurso) son categorías léxicas específicas a las que se asignan palabras en función de su **contexto sintáctico** y su **función**.
 - Con el contexto intentamos evitar la ambigüedad.
- Son útiles cuando necesitamos usar el mismo texto anotado más tarde en aplicaciones más complejas porque podemos filtrar por tags específicos y usar esa información para realizar análisis específicos.
- Por ejemplo:
 - Podemos reducir los sustantivos y determinar cuáles son los más destacados.
 - Podemos guardar solo los adjetivos para hacer un análisis de sentimientos rápido.
- La mayoría de los taggers se basan en la notación Penn Treebank.
- Dos categorías de palabras:
 - *Cerradas*. Tienen una categoría relativamente fija. Son palabras con “funciones claras” que tienden a ser cortas, aparecer frecuentemente y, generalmente, dan estructura al resto.
 - Por ejemplo, preposiciones.
 - *Abiertas*. Los conjuntos se modifican periódicamente.
 - Por ejemplo, sustantivos y verbos.

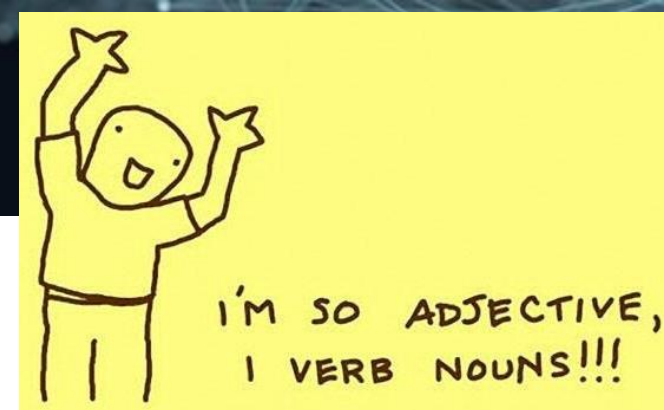


Análisis Sintáctico

Etiquetado Part-of-Speech

Diferentes técnicas:

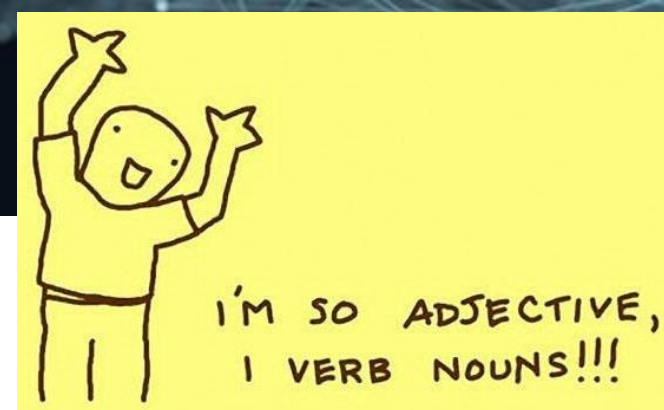
- **Léxicas.**
 - Se asigna el POS tag más común de una palabra en un corpus.
- **Reglas.**
 - Por ejemplo, “las palabras terminadas en -ed o -ing son verbos”.
 - Pueden ser utilizados en conjunto con los léxicos.
 - Para permitir etiquetar aquellas palabras que no aparecían en el training.
- **Probabilísticos.**
 - De acuerdo a la probabilidad de ocurrencia de una secuencia de tags.
 - Consideran el contexto.
 - Modelos de Markov ocultos o Conditional Random Fields.
- **Deep learning.**
 - Redes recurrentes.



Análisis Sintáctico

Etiquetado Part-of-Speech: basado en Reglas

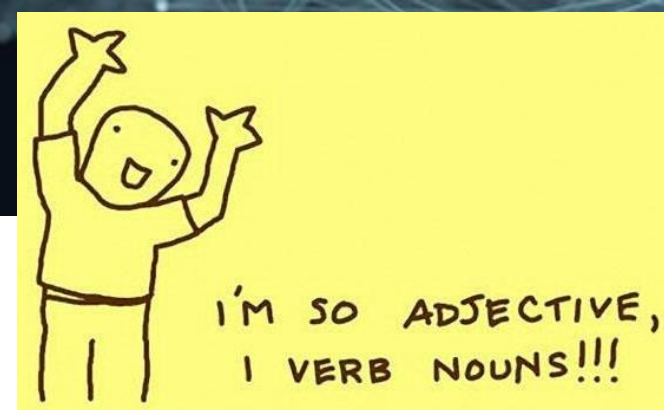
- Uno de los más antiguos.
- Usa las reglas para resolver las ambigüedades que surjan de un léxico.
 - Reglas definidas a mano.
 - Para el inglés, se han definido alrededor de 1000 reglas.
- Las reglas puede ser:
 - Expresiones regulares.
 - Patrones de contexto. (ej. "si la palabra que está antes de otra es un artículo, la siguiente debe ser un sustantivo.")
- Se pueden también inducir también a partir de un corpus (como el de Brown).
 - Se arranca con un léxico y se apunta a corregir los errores que puedan surgir.
 - Se define un conjunto de condiciones para determinar si se cambia o no el tag asignado a una palabra.



Análisis Sintáctico

Etiquetado Part-of-Speech: basado en Reglas

- Uno de los más antiguos.
- Usa las reglas para resolver las ambigüedades que surjan de un léxico.
 - Reglas definidas a mano.
 - Para el inglés, se han definido alrededor de 1000 reglas.
- Las reglas puede ser:
 - Expresiones regulares.
 - Patrones de contexto. (ej. "si la palabra que está antes de otra es un artículo, la siguiente debe ser un sustantivo.")
- Se pueden también inducir también a partir de un corpus (como el de Brown).



(1) TO IN NEXT-TAG AT

The first patch states that if a word is tagged **TO** and the following word is tagged **AT**, then switch the tag from **TO** to **IN**. This is because a noun phrase is much more likely to immediately follow a preposition than to immediately follow infinitive **TO**.

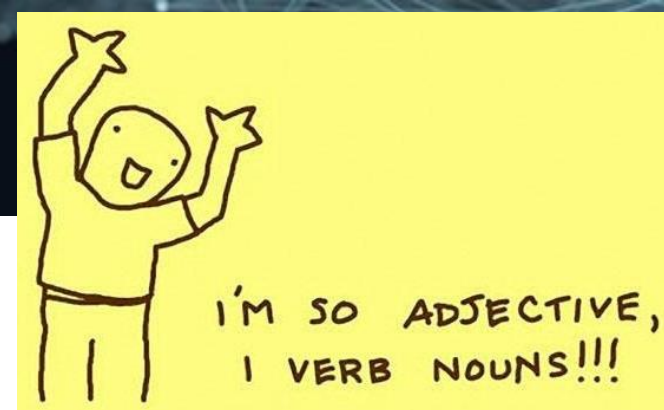
[A SIMPLE RULE-BASED PART OF SPEECH TAGGER](#)



Análisis Sintáctico

Etiquetado Part-of-Speech: probabilístico

- Se basan en la probabilidad de ocurrencia de un tag/palabra.
- Requiere un corpus de entrenamiento.
 - Y otro diferente para hacer el testing.
- Hay que controlar los casos en los que las palabras no existen en el corpus de entrenamiento.
- Basado en frecuencia de palabras.
 - Desambigua basándose en la probabilidad de que una palabra ocurra con un determinado tag.
 - Se asignará el tag más frecuente para la palabra en el corpus.
 - Puede dar lugar a secuencias imposibles de tags.
 - Para lenguajes fuertemente estructurados.
- Basado en la frecuencia de secuencias de tags.
 - Se basa en el cálculo de que ocurra una determinada secuencia de tags.



Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb



Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb

- Pueden estar acompañados de determinantes.
 - A goat, ITS bandwidth.
- Pueden aceptar posesivos.
 - Plato'S Republic.
- La matoría acepta formas plurales.
 - HouseS, indexES

Sustantivos

Propios

- No suelen estar precedidos por artículos o determinantes.
- En escrito, se suelen indicar con mayúscula.

Comunes

Contables

- Pueden ser contados.
- Tienen singular y plural.
- Singulares con artículo

Incontables

- No tienen cantidad.

Colectivos

- Representan un grupo homogéneo.
- Pueden no llevar artículo.



Noun

Verb

ADJective

ADVerb

- Se refieren a las acciones y procesos.
- Pueden tener inflexiones.
 - Personas.
 - Género.
 - Gerundio.
 - Participio.
- Se ha discutido si todos los idiomas realizan la distinción entre verbos y sustantivos.
 - Por ejemplo, Riau Indonesian y Tongan no la realizan.

Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb

- Propiedades o cualidades.
- Descripciones.
- Hay idiomas que no tienen adjetivos.
 - Koreano. Lo que en Inglés son adjetivos, acá son sub-clases de verbos.
 - Por ejemplo, “beautiful” actúa como un verbo “to be beautiful”.
- Idiomas donde los adjetivos se “confunden” con los sustantivos.
 - Georgian. Los sustantivos pueden modificar a otros.
- Idiomas que aplican inflexiones.
 - Género y número.
 - Tiempo verbal. → Mohawk



Noun

Verb

ADJective

ADVerb

- Principalmente modifican verbos.
 - También pueden modificar otros adverbios o frase verbales.
- Distintos tipos:
 - Direccionales. Especifican la dirección o ubicación de alguna acción.
 - De grado. El “nivel” o “alcance” de alguna acción, proceso o propiedad.
 - Forma. La manera en que se realiza una acción o proceso.
 - Temporales. El momento en el que se realiza la acción o proceso.
 - Pueden ser confundidos con sustantivos.



Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb

Tipos
cerrados

Difieren de lenguaje a lenguaje.

Preposiciones

on, under, over, near, by, at, from, to, with

- Ocurren antes de los sustantivos.
- Semánticamente generalmente presentan relaciones espaciales o temporales (literales o metafóricas).
- Pueden presentar relaciones indicando agentes.

Conjunciones

and, but, or, as, if, when

- Unen dos frases, cláusulas u oraciones.
- Coordinantes. Unen dos elementos del mismo status.
- Subordinantes. Uno de los elementos tiene un status subordinado a la otra.
 - Por ejemplo, una hace de contenido de la otra.



Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb

Tipos
cerrados

Difieren de lenguaje a lenguaje.

Partículas

up, down, on, off, in, out...

- Parecen preposiciones o adverbios.
- Son utilizadas en combinación con verbos.
- Tienen significados que pueden ser distintos que aquellos de las preposiciones a las que se parecen.
 - Por ejemplo, 'she turned the paper over'.

Determinantes

a, an, the, this, that

- Los artículos son un sub-conjunto.
- Definidos e indefinidos.
- "The" es la palabra más frecuente en el idioma Inglés.
- No todos los idiomas tienen artículos.
- Árabe no tiene indefinidos.
- Rumano tiene enclítico.



Análisis Sintáctico

Etiquetado Part-of-Speech

Noun

Verb

ADJective

ADVerb

Tipos
cerrados

Difieren de lenguaje a lenguaje.

Pronombres

she, who, I, what, who, ...

- Formas "abreviadas" para referirse a alguna frase sustantiva, entidad o evento.
- Pronombres personales se refieren a personas o entidades.
- Pronombres posesivos indican posesión o una relación abstracta entre la persona y un objeto.
- Wh-pronouns preguntas o en subordinaciones.

Verbos modales y auxiliares

can, may, should
be, have, do

- Auxiliares marcan aspectos semánticos de los verbos principales.
- Modales marcan la forma asociada con el evento del verbo principal.
- Can. Habilidad o posibilidad.
- May. Permiso o posibilidad.
- Must. Obligación.



Análisis Sintáctico

Etiquetado Part-of-Speech

SI No.	TAG	DESCRIPTION	EXAMPLE(S)
1	CC	Coordinating conjunction	and, or
2	CD	Cardinal number	five, one, 2
3	DT	Determiner	a, the
4	EX	Existential <i>there</i>	there were two cars
5	FW	Foreign word	d'hoevre, mais
6	IN	Preposition/subordinating conjunction	of, in, on, that
7	JJ	Adjective	quick, lazy
8	JJR	Adjective, comparative	quicker, lazier
9	JJS	Adjective, superlative	quickest, laziest
10	LS	List item marker	2)
11	MD	Verb, modal	could, should
12	NN	Noun, singular or mass	fox, dog
13	NNS	Noun, plural	foxes, dogs
14	NNP	Noun, proper singular	John, Alice
15	NNPS	Noun, proper plural	Vikings, Indians, Germans
16	PDT	Predeterminer	both cats
17	POS	Possessive ending	boss's
18	PRP	Pronoun, personal	me, you
19	PRP\$	Pronoun, possessive	our, my, your
20	RB	Adverb	naturally, extremely, hardly
21	RBR	Adverb, comparative	better
22	RBS	Adverb, superlative	best
23	RP	Adverb, particle	about, up
24	SYM	Symbol	%, \$
25	TO	Infinitival to	how to, what to do

SI No.	TAG	DESCRIPTION	EXAMPLE(S)
26	UH	Interjection	oh, gosh, wow
27	VB	Verb, base form	run, give
28	VBD	Verb, past tense	ran, gave
29	VBG	Verb, gerund/present participle	running, giving
30	VBN	Verb, past participle	given
31	VBP	Verb, non-third person singular present	I think, I take
32	VBZ	Verb, third person singular present	he thinks, he takes
33	WDT	Wh-determiner	which, whatever
34	WP	Wh-pronoun, personal	who, what
35	WP\$	Wh-pronoun, possessive	whose
36	WRB	Wh-adverb	where, when
37	NP	Noun phrase	the brown fox
38	PP	Prepositional phrase	in between, over the dog
39	VP	Verb phrase	was jumping
40	ADJP	Adjective phrase	warm and snug
41	ADVP	Adverb phrase	also
42	SBAR	Subordinating conjunction	whether or not
43	PRT	Particle	up
44	INTJ	Interjection	hello
45	PNP	Prepositional noun phrase	over the dog, as of today
46	-SBJ	Sentence subject	the fox jumped over the dog
47	-OBJ	Sentence object	the fox jumped over the dog



- Se debe poder tratar con palabras desconocidas.
 - Sustantivos propios y acrónimos son creados a menudo.
 - Sustantivos comunes y verbos también se modifican.
- Una forma de realizar las distinciones es considerar la morfología de las palabras.
 - Sustantivos propios suelen comenzar con mayúscula.
 - Palabras finalizadas en -s tienden a ser sustantivos plurales.
 - Palabras finalizadas en -ed tienden a ser participios.
 - Palabras finalizadas en -able adjetivos.
 - ...
- Otros idiomas proveen más información que el Inglés en las inflexiones de las palabras.
 - Caso (Nominativo, acusativo, genitivo...)
 - Género
 - ...



- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.



- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

We saw the yellow dog

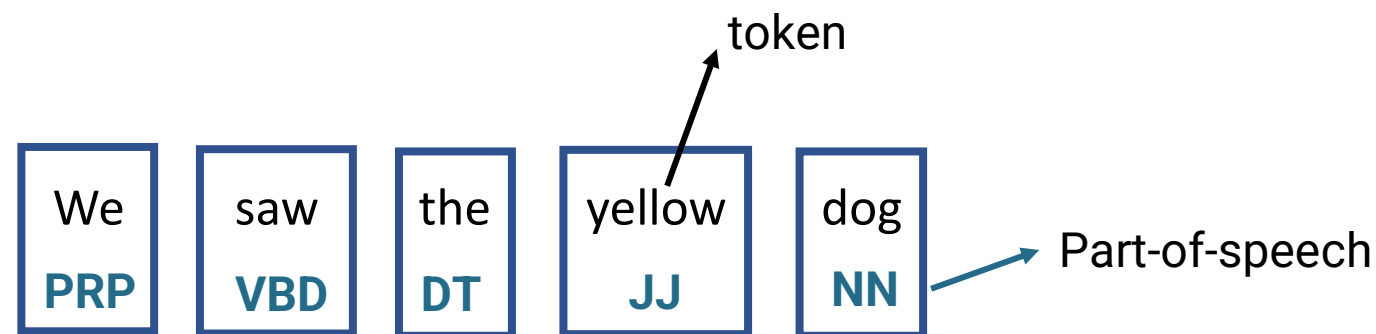


Análisis Sintáctico

Shallow parsing - Chunking

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

- PRP: Pronombre personal
- VBD: verbo en pasado
- DT: determinante
- JJ: adjetivo
- NN: sustantivo singular o colectivo

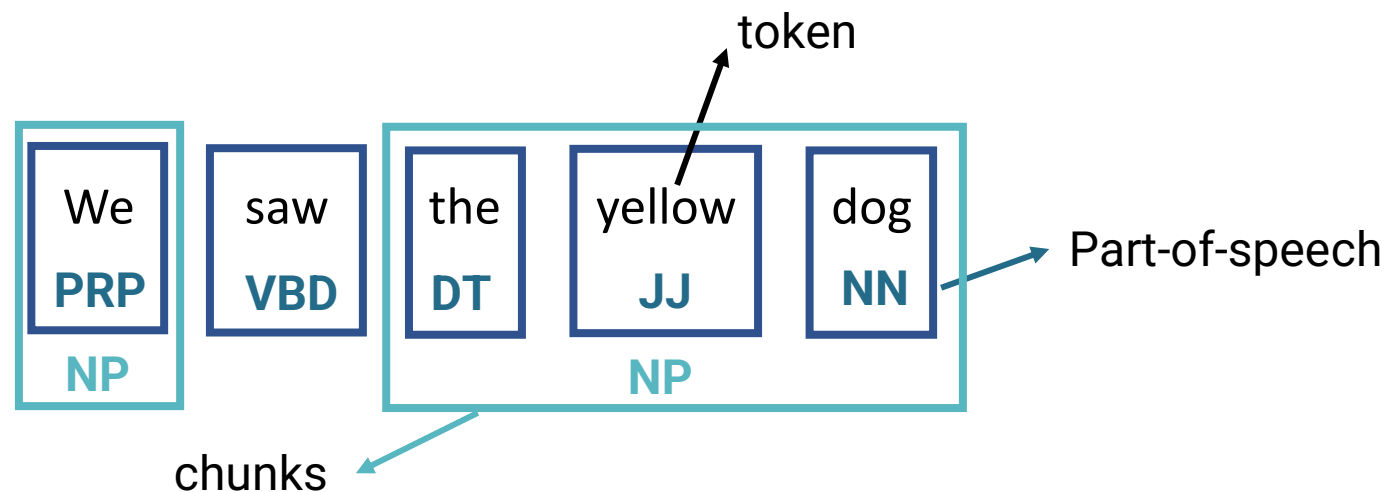


Análisis Sintáctico

Shallow parsing - Chunking

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

- PRP: Pronombre personal
- VBD: verbo en pasado
- DT: determinante
- JJ: adjetivo
- NN: sustantivo singular o colectivo



Análisis Sintáctico

Shallow parsing - Chunking

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

Noun Phrase

Verb Phrase

Adjective
Phrase

Adverb
Phrase

Prepositional
Phrase



- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

Noun Phrase

- Frases donde el sustantivo es la palabra principal.
- Actúan como el sujeto o el objeto de un verbo.
- Pueden comenzar con determinantes, o expresiones más complejas que los reemplacen.
 - Por ejemplo, una expresión posesiva.

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

Verb Phrase

- Frases donde el verbo es la palabra principal.
- Dos formas (asociadas a distintos tipos de parsers):
 - Contiene el verbo y otras entidades (sustantivos, adjetivos, adverbios) que son parte del objeto. El verbo actúa como una unidad única y funciona como la raíz.
 - Contiene solo componentes verbales, incluyendo verbo principal, auxiliaries, infinitivos y participios.
- No todos los verbos son compatibles con todas las posibles combinaciones de constituyentes.

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

Adjective Phrase

- Frases donde el adjetivo es la palabra principal.
- Su rol principal es describir sustantivos y pronombres.
- Se ubican antes o después de un sustantivo o pronombre.

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
 - El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
 - No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.
- Frases donde el adverbio es la palabra principal.
 - Se usan como modificadores para sustantivos, verbos u otros adverbios, proveyendo detalles que los describen o califican.

Adverb
Phrase

- El shallow parsing es una técnica de análisis de la estructura de una oración que agrupa los tokens en frases de un nivel superior de abstracción.
- El objetivo principal del análisis superficial es obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento, como si hacen otros tipos de parsing.

- Frases donde una preposición es la palabra principal.
- Actúan como un adjetivo o adverbio, describiendo otras palabras o frases.

Prepositional
Phrase



- La gramática ayuda a definir la sintaxis y la estructura en el lenguaje.
- Principalmente consiste en un conjunto de reglas que se utilizan para determinar cómo posicionar palabras, frases y cláusulas al construir oraciones en un lenguaje natural.
- La gramática se puede subdividir en dos clases principales en función de sus representaciones de sintaxis y estructura lingüísticas.
- “Parsing” se refiere a la tarea de crear un árbol de parsing para una oración determinada.
 - Resalta la estructura sintáctica de acuerdo a la gramática subyacente.

- Los árboles de parsing sirven como una importante etapa intermedia de representación para el análisis semántico. Importante para:
 - Responder preguntas.
 - Extracción de información.
- Problema: Ambigüedad.
 - Oraciones gramáticamente correctas pueden dar lugar a parsings sin sentido semántico.
 - Los parsers deben poder elegir cuál es el parsing correcto de todos los posibles.
 - Requiere conocimiento del contexto.



- Los árboles de parsing sirven como una importante etapa intermedia de representación para el análisis semántico. Importante para:
 - Responder preguntas.
 - Extracción de información.
- Problema: Ambigüedad.
 - Oraciones gramáticamente correctas pueden dar lugar a parsings sin sentido semántico.
 - Los parsers deben poder elegir cuál es el parsing correcto de todos los posibles.
 - Requiere conocimiento del contexto.

"One morning, I shot an elephant in my pajamas.

—Groucho Marx



- Los árboles de parsing sirven como una importante etapa intermedia de representación para el análisis semántico. Importante para:
 - Responder preguntas.
 - Extracción de información.
- Problema: Ambigüedad.
 - Oraciones gramáticamente correctas pueden dar lugar a parsings sin sentido semántico.
 - Los parsers deben poder elegir cuál es el parsing correcto de todos los posibles.
 - Requiere conocimiento del contexto.

"One morning, I shot an elephant in my pajamas. How he got in my pajamas I don't know."

—Groucho Marx



- Los árboles de parsing sirven como una importante etapa intermedia de representación para el análisis semántico. Importante para:
 - Responder preguntas.
 - Extracción de información.
- Problema: Ambigüedad.
 - Oraciones gramáticamente correctas pueden dar lugar a parsings sin sentido semántico.
 - Los parsers deben poder elegir cuál es el parsing correcto de todos los posibles.
 - Requiere conocimiento del contexto.

Constituency
Parser

Dependency
Parser



- Analiza la sintaxis y las reglas que rigen la jerarquía y el orden de los diversos componentes en las oraciones.
- Se basa en la identificación de grupos de palabras que se comportan como un conjunto, los **constituyentes**.
- Cada palabra generalmente pertenece a una categoría léxica específica en el caso y forma la palabra principal de diferentes frases.
- Estas frases se forman en base a reglas llamadas reglas de estructura de frases.
- Se basa en las **gramáticas libres del contexto**.



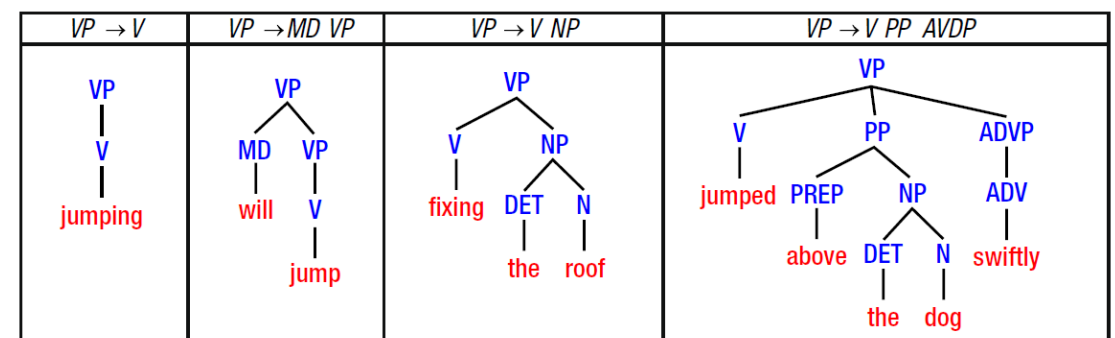
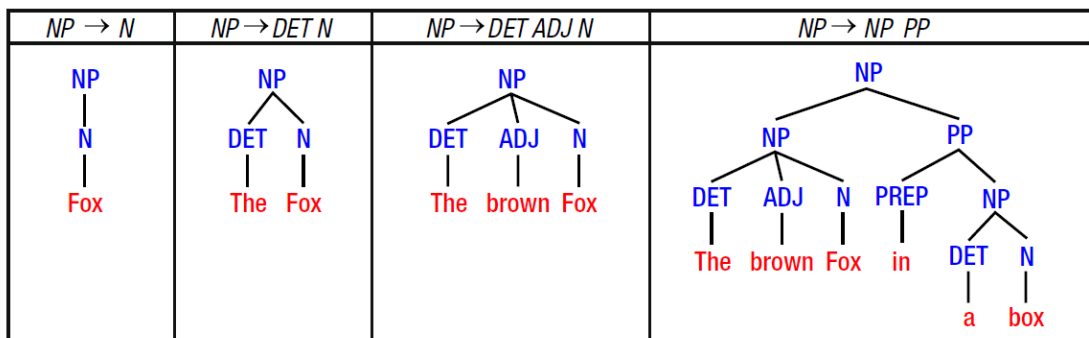
Constituency parsing: Gramáticas libres del contexto

- Una gramática consiste en un conjunto de reglas o producciones.
- Cada regla expresa la manera en la que los símbolos del lenguaje pueden ser agrupados y ordenados juntos.
- Dos tipos de símbolos.
 - *Terminales*. Se corresponden con las palabras del lenguaje.
 - *No terminales*. Representan abstracciones sobre los terminals. En este caso, las categorías POS y las estructuras de frases.
- Las reglas pueden ser vistas como generadoras en las que “el símbolo de la izquierda se reescribe como el símbolo de la derecha”.



Constituency parsing: Gramáticas libres del contexto

- Una gramática consiste en un conjunto de reglas o producciones.
- Cada regla expresa la manera en la que los símbolos del lenguaje pueden ser agrupados y ordenados juntos.
- Dos tipos de símbolos.
 - *Terminales*. Se corresponden con las palabras del lenguaje.
 - *No terminales*. Representan abstracciones sobre los terminals. En este caso, las categorías POS y las estructuras de frases.
- Las reglas pueden ser vistas como generadoras en las que “el símbolo de la izquierda se reescribe como el símbolo de la derecha”.



Constituency parsing

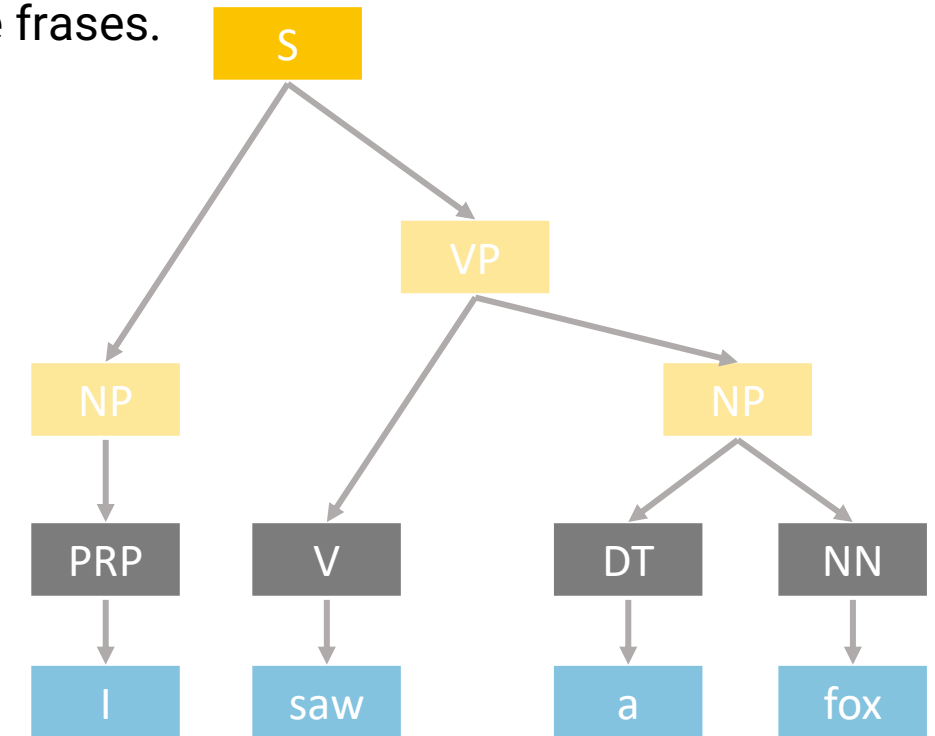
- Un árbol de constituency parsing siempre tiene las palabras de la oración como los nodos terminales.
- Cada palabra tiene un nodo padre que contiene su etiqueta POS.
 - En algunas representaciones gráficas puede omitirse.
- El resto de los nodos no terminales los representan las estructuras de frases.



Análisis Sintáctico

Constituency parsing

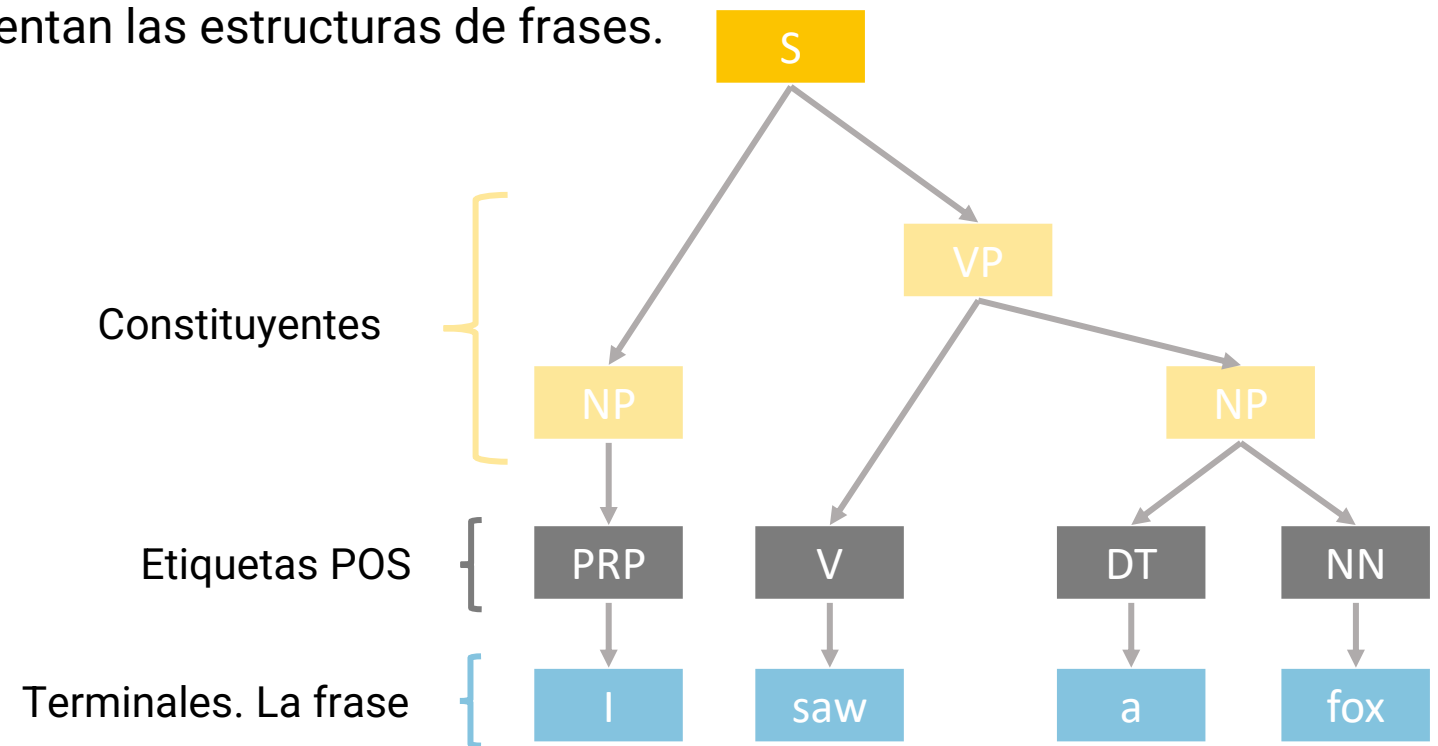
- Un árbol de constituency parsing siempre tiene las palabras de la oración como los nodos terminales.
- Cada palabra tiene un nodo padre que contiene su etiqueta POS.
 - En algunas representaciones gráficas puede omitirse.
- El resto de los nodos no terminales los representan las estructuras de frases.



Análisis Sintáctico

Constituency parsing

- Un árbol de constituency parsing siempre tiene las palabras de la oración como los nodos terminales.
- Cada palabra tiene un nodo padre que contiene su etiqueta POS.
 - En algunas representaciones gráficas puede omitirse.
- El resto de los nodos no terminales los representan las estructuras de frases.



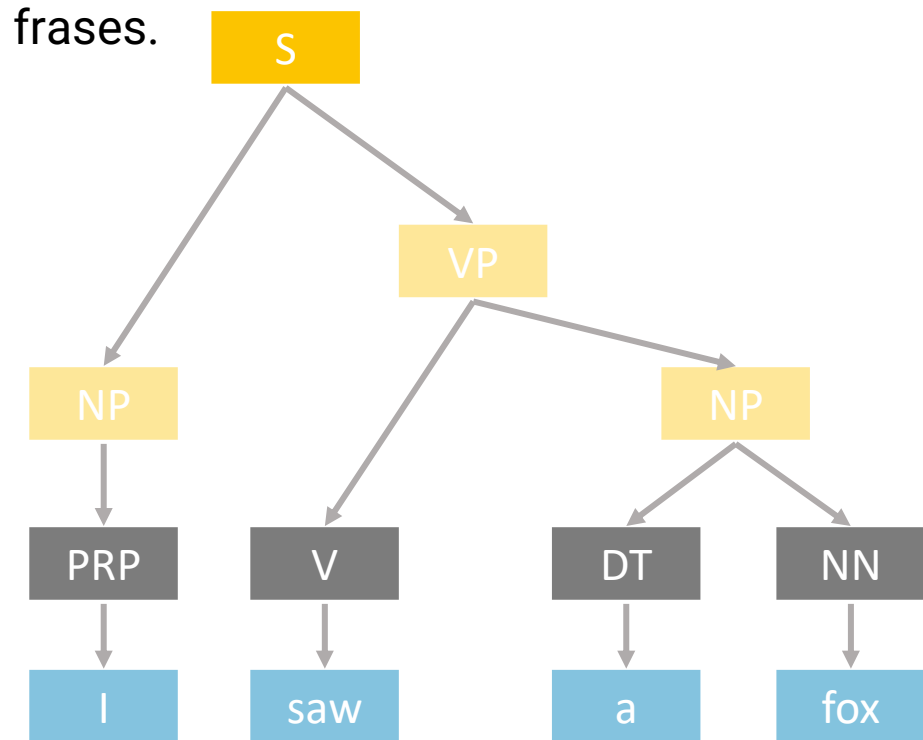
Análisis Sintáctico

Constituency parsing

- Un árbol de constituency parsing siempre tiene las palabras de la oración como los nodos terminales.
- Cada palabra tiene un nodo padre que contiene su etiqueta POS.
 - En algunas representaciones gráficas puede omitirse.
- El resto de los nodos no terminales los representan las estructuras de frases.
- En el primer nivel debajo de la raíz tenemos la división entre una frase nominal (NP) y una frase verbal (VP).
- Esto significa que la gramática tiene una regla que permite que una oración se componga con la concatenación de una NP y una VP.

$S \rightarrow NP VP$

- Luego, la frase verbal es dividida entre el verbo y otra frase nominal.
- Todas las relaciones se establecen mediante reglas!
- Si no hay regla, no se puede hacer la traducción.



- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.

- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
 - Una oración contiene una frase preposicional (PP) luego de un verbo con un complemento, haciendo ambigua la determinación de a qué asociar la frase preposicional.

- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
 - Una oración contiene una frase preposicional (PP) luego de un verbo con un complemento, hacienda ambigua la determinación de a qué asociar la frase preposicional.

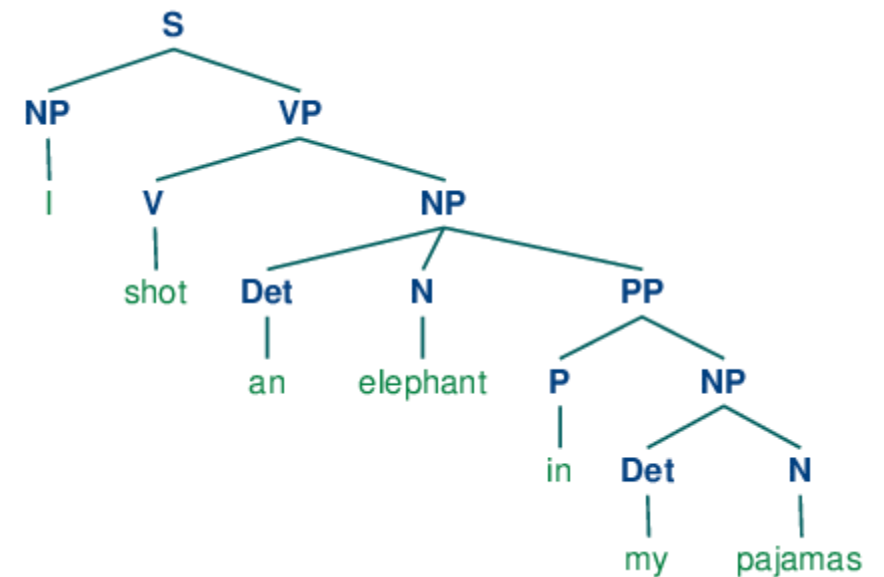
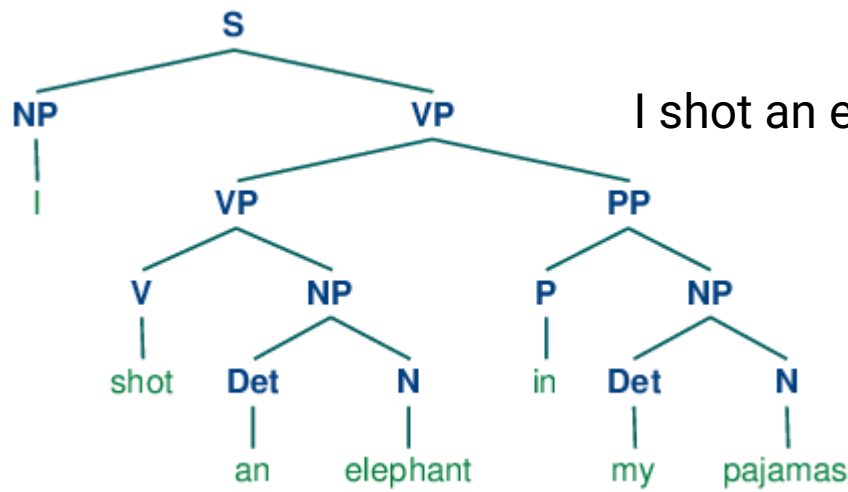
I shot an elephant in my pajamas



Análisis Sintáctico

Constituency parsing

- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
 - Una oración contiene una frase preposicional (PP) luego de un verbo con un complemento, hacienda ambigua la determinación de a qué asociar la frase preposicional.



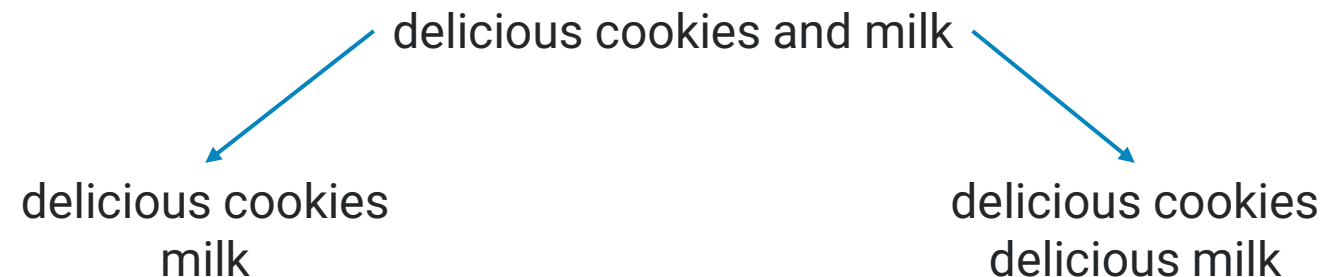
- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.



Análisis Sintáctico

Constituency parsing

- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.



- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.

The thief shot the hostage *and the cop*.



- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.

The thief shot the hostage *and the cop*.

The thief shot the hostage *and the cop* during a robbery.

The thief shot the hostage *and the cop* panicked.



- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.

The thief shot the hostage *and the cop*.

The thief shot **the hostage and the cop** during a robbery.

The thief shot the hostage *and the cop* panicked.



Constituency parsing

- Uno de los problemas más “serios” es el de la ambigüedad.
- Ambigüedad estructural.
- Dos formas.
- **Attachment ambiguity.** Un constituyente puede ser agregado al árbol en más de un lugar.
- **Coordination ambiguity.** Las conjunciones dentro de una oración tienen múltiples interpretaciones.
 - El modificador puede aplicarse a toda la conjunción o a un único componente de la misma.
 - Generalmente, es el caso de una combinación incluyendo sustantivos o frases nominales y un adjetivo o frase adjetiva.

The thief shot the hostage *and the cop*.

The thief shot **the hostage and the cop** during a robbery.

The thief shot the hostage *and the cop panicked*.

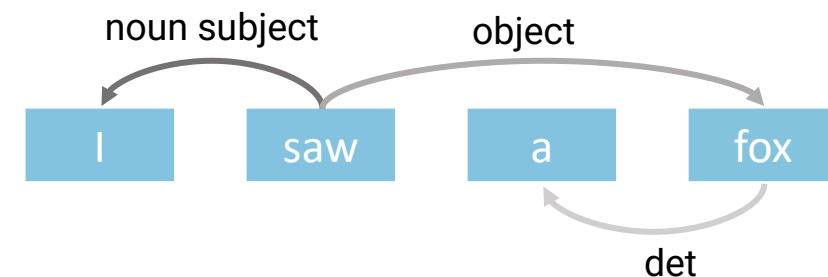


- A diferencia del constituency, no hace uso de las frases constituyentes o de las sub-frases.
- La sintaxis es expresada en términos de las dependencias directas entre palabras en un grafo.
- Gramáticas basadas en la dependencia analizan e infieren tanto la estructura como las dependencias semánticas y las relaciones entre los tokens en una oración.
- En cualquier oración del idioma, todas las palabras, excepto una, tienen alguna relación o dependencia a otras palabras en la oración.
 - Relaciones son binarias.
 - Proveen una indicación de la semántica de la relación.
- La palabra que no tiene dependencia se llama raíz de la oración.
 - El **verbo** se toma como la raíz de la oración en la mayoría de los casos.
 - Todas las otras palabras están directa o indirectamente vinculadas al verbo raíz mediante enlaces, que son las dependencias.



- Un árbol de dependency parsing es un grafo $G = (V, E)$ donde:
 - V : los nodos son las palabras en la oración.
 - E : los arcos conectan dos palabras y tienen asociado un tipo, el cual define la relación gramática entre las palabras.
- El grafo debe satisfacer tres condiciones.
 - Debe haber un único nodo raíz sin arcos entrantes.
 - Cada nodo excepto la raíz debe tener un único arco entrante.
 - Hay un camino único entre la raíz y cada uno de los nodos.
- Estas condiciones garantizan que:
 - Cada palabra tiene una única “cabeza”.
 - La estructura del grafo está conectada.

- Un árbol de dependency parsing es un grafo $G = (V, E)$ donde:
 - V : los nodos son las palabras en la oración.
 - E : los arcos conectan dos palabras y tienen asociado un tipo, el cual define la relación gramática entre las palabras.
- El grafo debe satisfacer tres condiciones.
 - Debe haber un único nodo raíz sin arcos entrantes.
 - Cada nodo excepto la raíz debe tener un único arco entrante.
 - Hay un camino único entre la raíz y cada uno de los nodos.
- Estas condiciones garantizan que:
 - Cada palabra tiene una única “cabeza”.
 - La estructura del grafo está conectada.



- Permiten clasificar los tipos de relaciones o funciones gramaticales.
 - Las más conocidas: sujeto, objeto directo, objeto indirecto.
- Dos tipos de relaciones:
 - Clausal. Describen los roles sintácticos con respecto a un predicado.
 - Modificadores. Categoriza las maneras en la que una palabra puede modificar a su cabeza.
- Dada la estructura del idioma, en Inglés, estas funciones se encuentran relacionadas (pero no determinadas) con la posición en la oración y la frase constituyente.
- No sucede lo mismo en lenguajes menos estructurados sin un orden en las palabras tan claro.
 - En este tipo de lenguajes, los constituency parsers necesitarían una regla separada para cada uno de los casos a contemplar.
 - Por el contrario, un dependency parser solo necesita tener un vínculo expresando esta relación, abstrayéndose del orden.
- Se utilizan métodos supervisados creados en base a corpus (Penn Treebank corpus otra vez).

Constituency parser

- Divide las oraciones en sub-frases.
- Los no terminales en el árbol de parsing son los tipos de frases.
- Los terminales son las palabras en la oración.
- Los arcos no tienen etiquetas.
- Permite visualizar la estructura completa de la oración y no solo las dependencias gramaticales.

Dependency parser

- Conecta las palabras de acuerdo a las relaciones entre ellas.
- Cada vértice en el árbol representa una palabra.
- Los nodos hijo son palabras que dependen del padre.
- Los arcos están etiquetados con la relación.



Constituency parser

- Divide las oraciones en sub-frases.
- Los no terminales en el árbol de parsing son los tipos de frases.
- Los terminales son las palabras en la oración.
- Los arcos no tienen etiquetas.
- Permite visualizar la estructura completa de la oración y no solo las dependencias gramaticales.

Dependency parser

- Conecta las palabras de acuerdo a las relaciones entre ellas.
- Cada vértice en el árbol representa una palabra.
- Los nodos hijo son palabras que dependen del padre.
- Los arcos están etiquetados con la relación.

Mary calls Peter

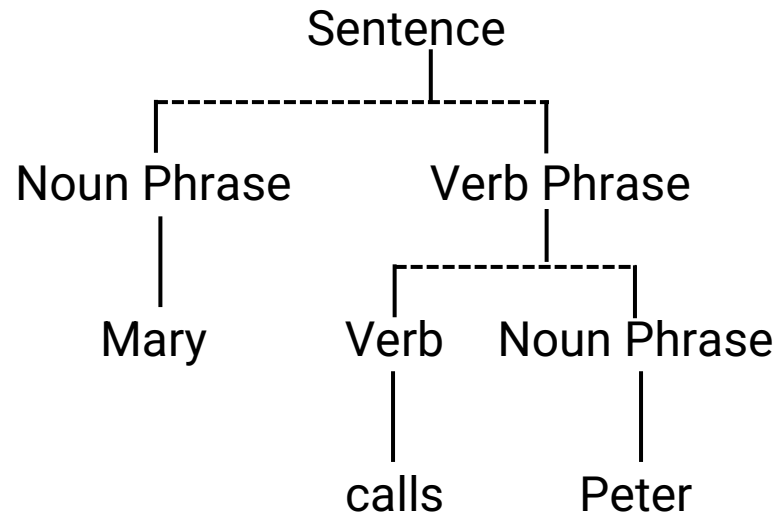


Análisis Sintáctico

Constituency parsing vs Dependency parsing

Constituency parser

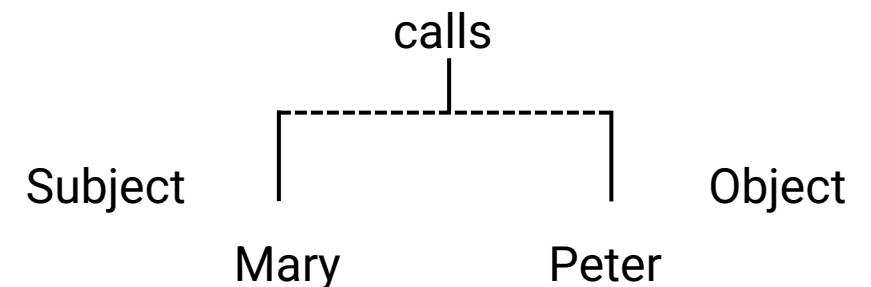
- Divide las oraciones en sub-frases.
- Los no terminales en el árbol de parsing son los tipos de frases.
- Los terminales son las palabras en la oración.
- Los arcos no tienen etiquetas.
- Permite visualizar la estructura completa de la oración y no solo las dependencias gramaticales.



Mary calls Peter

Dependency parser

- Conecta las palabras de acuerdo a las relaciones entre ellas.
- Cada vértice en el árbol representa una palabra.
- Los nodos hijo son palabras que dependen del padre.
- Los arcos están etiquetados con la relación.

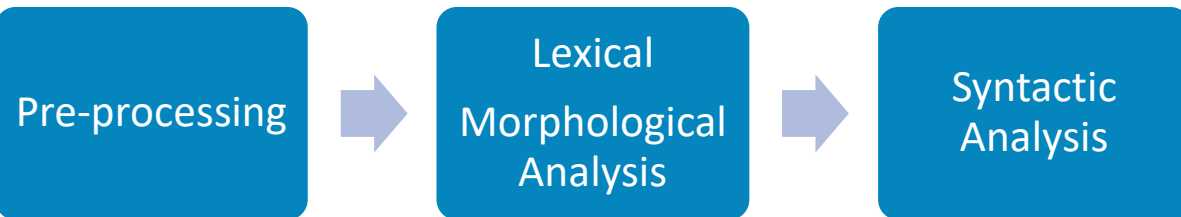


Constituency parsing vs Dependency parsing:Cuál utilizar?

- Depende de para qué se vayan a utilizar las estructuras detectadas.
- Ambos parsers se pueden utilizar para la extracción de características para entrenar modelos supervisados.
- **Dependency parsing.**
 - Útil para tareas como Extracción de Información o Question Answering.
 - Sencillo extraer las triplas sujeto-verbo-objetos que son a menudo indicadoras de las relaciones semánticas entre los predicados.
 - Para lenguajes que no imponen un orden específico a las palabras en una oración.
- **Constituency parsing.**
 - Extraer sub-frases de una oración.
 - No es que no se puedan realizar las tareas que se describen para Dependency, sino que hay que agregar procesamiento al árbol.
 - Chequeo de gramática.



Resumen!



- **Análisis Sintáctico.** La sintaxis y la estructura de las oraciones generalmente están relacionadas, y existe un conjunto de reglas, convenciones y principios específicos que rigen la forma en que se combinan las palabras.

Part-of-speech

- Categorías léxicas específicas a las que se asignan palabras en función de su **contexto sintáctico** y su **función**.
- Son útiles cuando necesitamos usar el mismo texto anotado más tarde en aplicaciones más complejas.

Constituency parsing

- Divide las oraciones en sub-frases.
- Los no terminales en el árbol de parsing son los tipos de frases.
- Los terminales son las palabras en la oración.
- Los arcos no tienen etiquetas.

Shallow parsing chunking

- Agrupa los tokens en frases de un nivel superior de abstracción.
- Obtener frases semánticamente significativas y observar las relaciones entre ellas.
- No profundiza en los detalles de la sintaxis interna ni todas las relaciones dentro de cada fragmento.

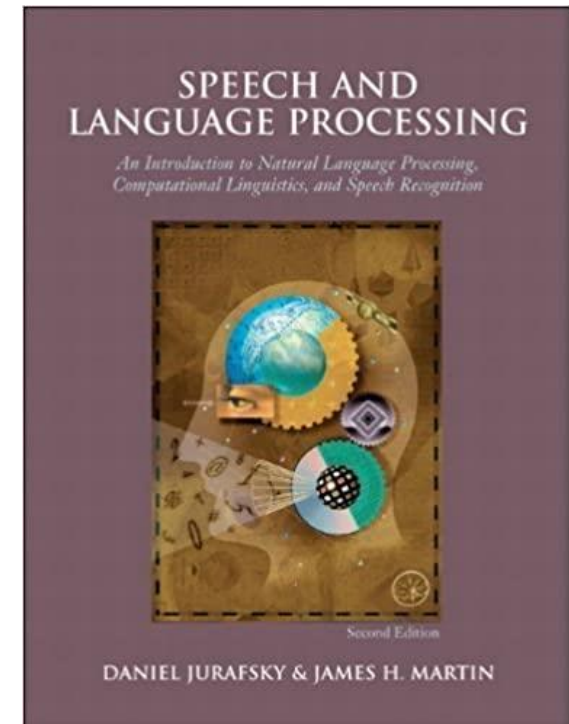
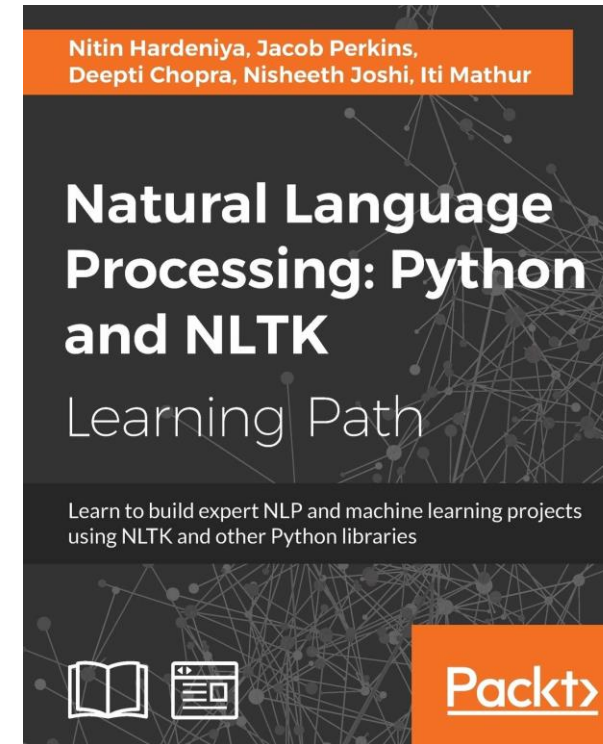
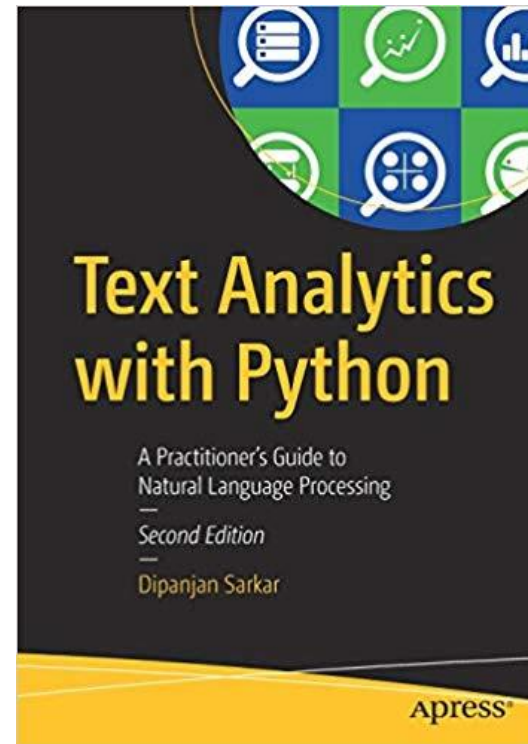
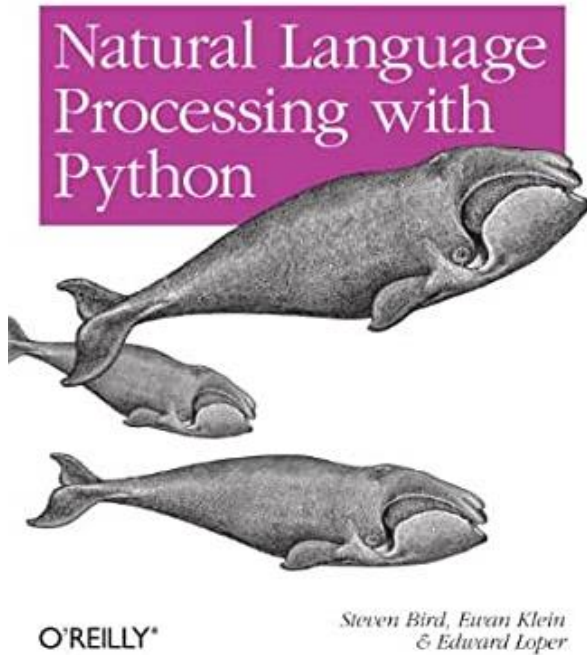
Dependency parsing

- Conecta las palabras de acuerdo a las relaciones entre ellas.
- Cada vértice en el árbol representa una palabra.
- Los nodos hijo son palabras que dependen del padre.
- Los arcos están etiquetados con la relación.



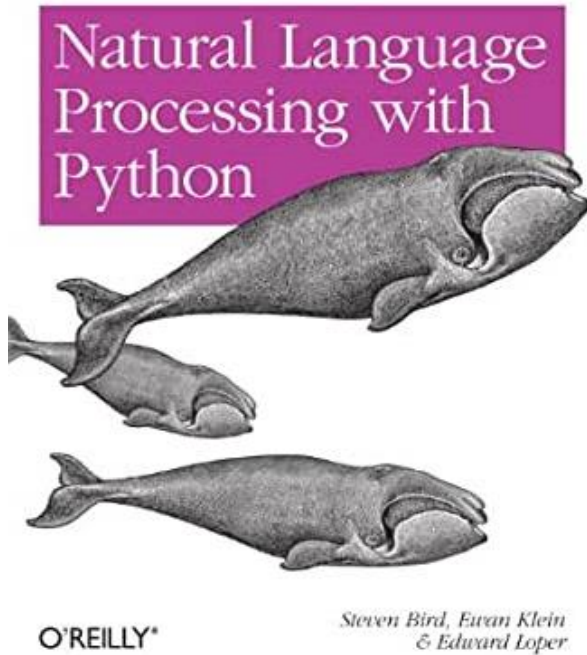
Lecturas sugeridas

Analyzing Text with the Natural Language Toolkit

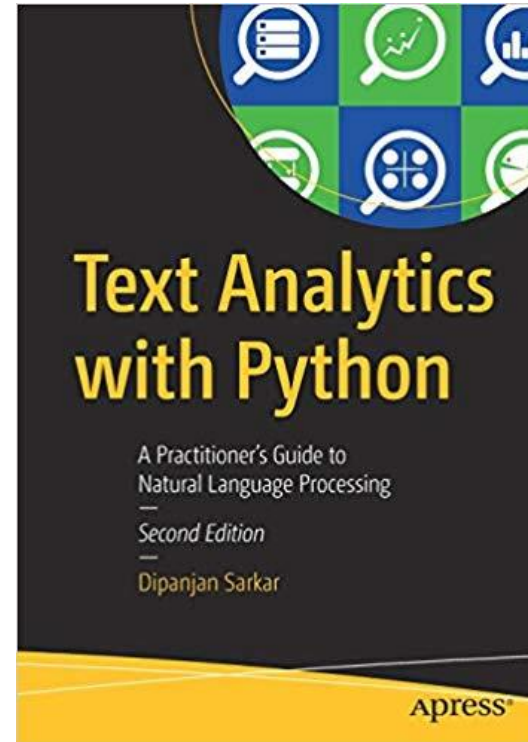


Lecturas sugeridas

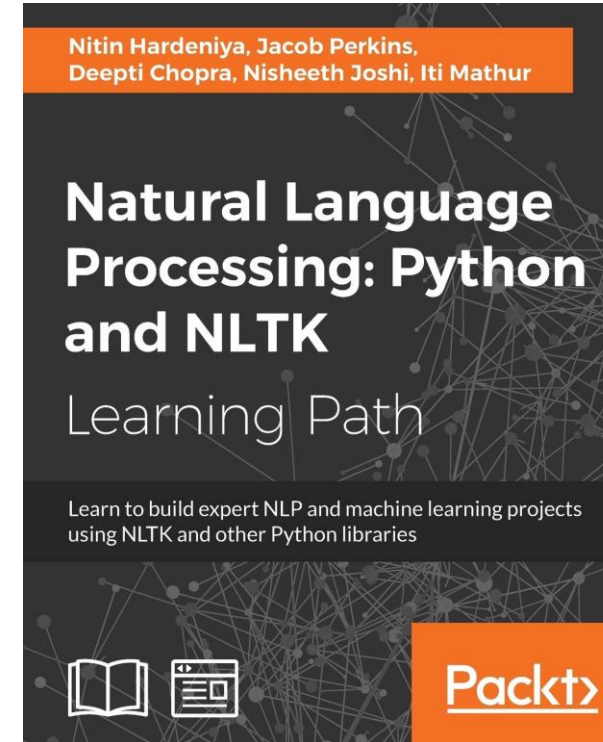
Analyzing Text with the Natural Language Toolkit



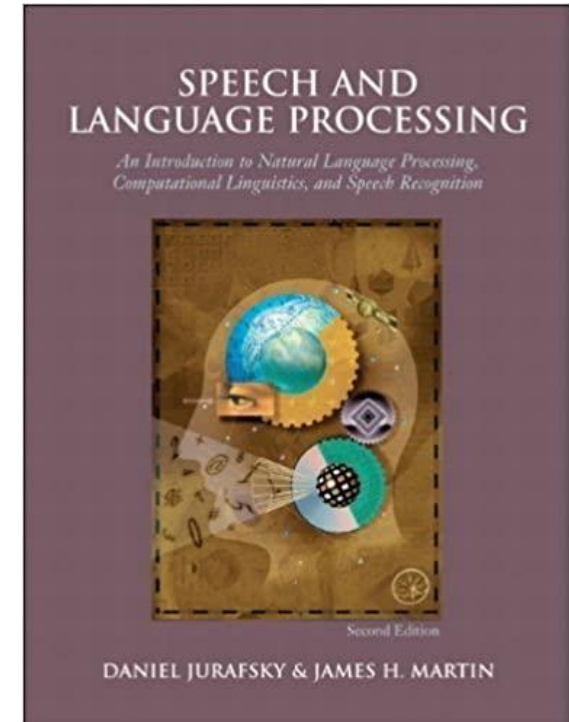
Capítulo 8. "Analyzing Sentence Structure"



Capítulo 1. "Natural Language Processing Basics"



Capítulo 3. "Part of Speech Tagging"
Capítulo 4. "Parsing Structure in Text"



Capítulo 13 (2nd Edition). "Syntactic Parsing"
Capítulo 14 (2nd Edition). "Statistical Parsing"
Capítulo 8 (3rd Edition). "Part-of-Speech Tagging"



diplomatura universitaria en
inteligencia artificial



FACULTAD DE CIENCIAS
EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES



diplomatura universitaria en
inteligencia artificial



FACULTAD DE CIENCIAS
EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO
DE LA PROVINCIA DE BUENOS AIRES

Procesamiento de Lenguaje Natural

Análisis Sintáctico