PYTHON PARA LINGÜISTAS

ANÁLISIS SINTÁCTICO

ALEJANDRO ARIZA

CENTRE DE LLENGUATGE I COMPUTACIÓ

UNIVERSITAT DE BARCELONA

¿QUÉ HEMOS VISTO?

- Fundamentos de programación:
  - Tipos de datos y variables
  - Funciones y métodos
  - Estructuras de decisión
  - Bucles
- Preprocesado básico de corpus:
  - Limpieza
  - Tokenización y segmentación de frases
  - Estadísticas simples de un corpus y n-gramas
  - Etiquetado POS. Aprendizaje automático supervisado.

SINTAXIS

- La sintaxis considera el orden y estructura de los elementos que aparecen en un texto.
- Cómo las palabras se combinan para expresar una idea.
- La sintaxis se aplica tanto a lenguajes computacionales como a lenguajes naturales.

  e.g.: en Python – cómo definir una función, qué parámetro corresponde a qué valor, etc
- Normalmente, pensamos en la sintaxis únicamente como el “orden de las palabras” pero también existe una jerarquía.

SINTAXIS (2)

- Preguntas frecuentes acerca de la sintaxis:
  - **Constituyentes**: ¿Cómo se agrupan las palabras?
  - **Relaciones gramáticas**: ¿Qué tipo de relaciones tienen estos grupos con respecto al verbo?
  - **Dependencias**: ¿Qué tipo de relaciones tienen las palabras individuales entre sí?

CONSTITUYENTES

- Grupos de palabras que se comportan como una única unidad o frase se conocen como constituyente:
  - **John** often comes late to class.
  - **My friend and I both** have a dog named Spot.
  - **Many parts of the Asian coastline** were destroyed by a tsunami in 2004.
  - **The old hotel at the end of the street** is going to be knocked down to make way for a new supermarket.
  - Sitting in a tree at the bottom of the garden was **a huge black bird with long blue tail feathers**.

ANÁLISIS DE CONSTITUYENTES

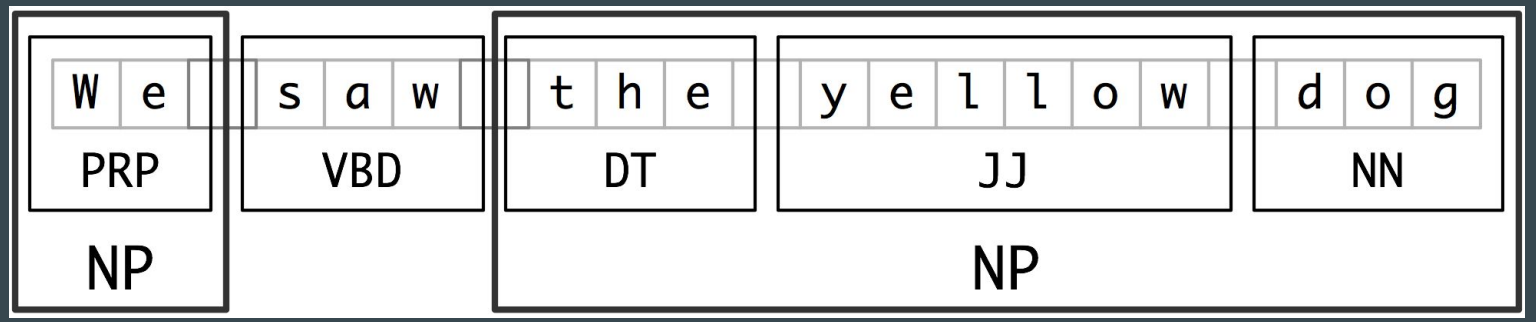
- El primer paso de un análisis sintáctico
- Identifica los constituyentes en una oración
- El análisis de constituyentes automático a veces recibe el nombre de “shallow parsing”
- Para muchas aplicaciones prácticas, un análisis de constituyentes contiene suficiente información y no necesitamos realizar un análisis sintáctico completo

RELACIONES GRAMATICALES

- Las relaciones gramaticales son relaciones funcionales entre constituyentes dentro de una oración
- Ejemplos comunes de relaciones gramaticales pueden ser el sujeto, objeto directo y objeto indirecto
- En NLP, las relaciones gramaticales se representan normalmente usando un Context Free Grammar (CFG)

I-O-B CHUNKING

- Un método simple de shallow parsing: **I**nside, **O**utside, **B**eginning
- Al marcar los tokens con estas 3 etiquetas (“inside”, “outside”, “beginning”), el I-O-B chunker identifica el comienzo y final de cada constituyente
- El I-O-B chunker normalmente solo utiliza las etiquetas POS e ignora las palabras



CONTEXT FREE GRAMMARS

- Un sistema matemático para modelar la estructura de constituyentes en inglés y otros idiomas
- Normalmente, tiene una visualización interpretable en términos de árboles jerárquicos
- La primera vez que se utilizó para análisis y descripción del lenguaje fue por N. Chomsky (1956)

CONTEXT FREE GRAMMAR. REGLAS

- NP -> Det Nominal — the flight
- NP -> ProperNoun — John
- Nominal -> Noun| Noun Nominal |Pronoun — flight, John, car, I, we, ...

- (Fijaros que Nominal es recursivo – contiene “Nominal” como posible expansión)

CONTEXT FREE GRAMMAR. LÉXICO

- Det -> a
- Det -> the
- Noun -> flight
- Noun -> car
- Pronoun -> I, we, you, they, ...

CONTEXT FREE GRAMMAR. REGLAS (2)

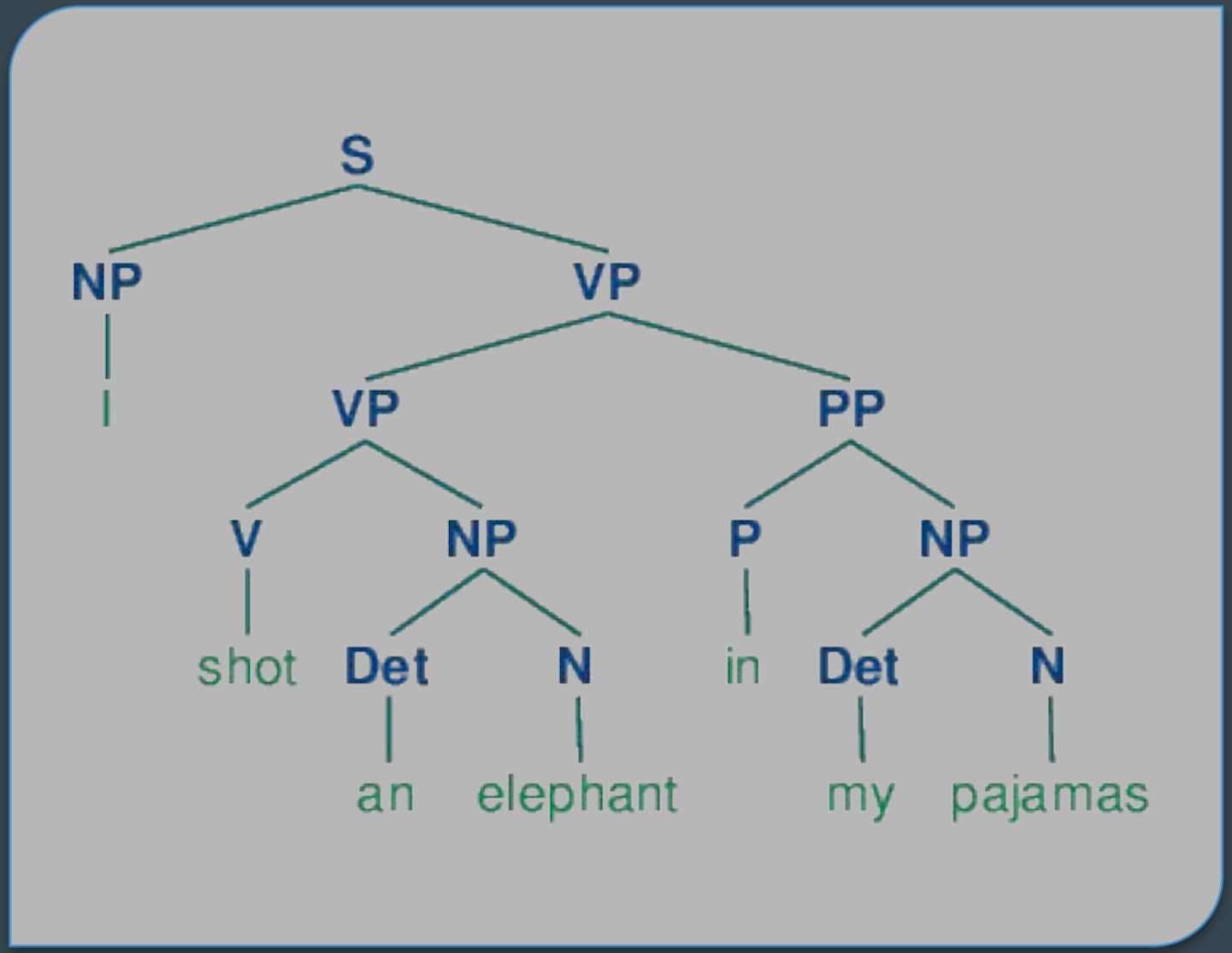
- En cada regla CF:
- Símbolo terminal: corresponde a palabras de un lenguaje (Léxico)
  - Det -> a
  - Det -> the
  - Noun -> flight
- Símbolo no terminal: Elementos del meta-lenguaje (generalizaciones)
  - NP -> Det Nominal
  - Nominal -> Noun| Noun Nominal

CONTEXT FREE GRAMMAR. REGLAS (3)

- En cada regla CF:
  - El símbolo a la izquierda: siempre un único símbolo no terminal
  - El símbolo a la derecha: una lista ordenada de uno o más símbolos terminales, o no terminales, o ambos
- En el léxico, el símbolo no terminal asociado a cada palabra es normalmente su etiqueta POS:
  - Det -> a

CONTEXT FREE GRAMMAR. REGLAS (4)

- Un CFG tiene 4 parámetros:
  - Un conjunto the símbolos no terminales, N (o variables: NP, V, PP, ...)
  - Un conjunto de símbolos terminales, K (diferente a N: a, the, John, cat, ...)
  - Un conjunto de reglas de producción P, de la forma A -> a, donde A es un símbolo no terminal y a es un string de símbolos contenidos en un conjunto infinito de strings (K *union* N)*
  - Un símbolo axiomático inicial S
- Un lenguaje se define a través del concepto de derivación

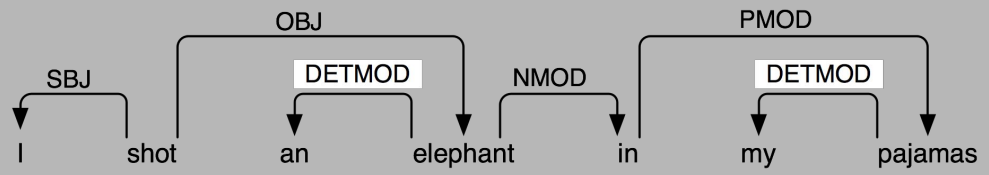


REGLAS CFG Y ÁRBOLES

- Una oración representada con un CFG puede visualizarse como un árbol jerárquico

CFG. GENERACIÓN Y ANÁLISIS

- Una CFG puede verse como una herramienta para generar frases:
  - Dado un conjunto de reglas y un léxico – genera una posible frase
- Un CFG puede ser visto también como una herramienta para asignar una estructura a una frase específica:
  - ¿Puede esta frase ser derivada usando una gramática predefinida?
  - ¿Qué tipo de constituyentes tiene una frase?
  - ¿Cuáles son las relaciones gramaticales en una frase?



DEPENDENCIAS

- Las dependencias son similares a categorías gramaticales. Sin embargo, las dependencias conectan palabras en vez de constituyentes

ANALIZADOR SINTÁCTICO

- Una herramienta (o programa) que dado un texto y una gramática puede:
  - Decir si el texto puede ser generado a partir de la gramática
  - Asignar una estructura al texto de acuerdo a la gramática
- Analizador Top-down: Busca un árbol sintáctico intentando construirlo desde el nodo raíz S hasta las “hojas” (palabras)
- Analizador Bottom-up: comienza por las palabras del texto e intenta construir el árbol sintáctico hacia el nodo raíz S

TOP-DOWN VERSUS BOTTOM-UP

- Top down:
  - Nunca gasta tiempo explorando árboles que no pueden terminar en el nodo raíz S
  - Nunca explora sub-árboles que no tienen cabida en un árbol con raíz S
  - Requiere un intensivo esfuerzo al analizar árboles que no encajan con la oración
- Bottom-up
  - Genera árboles que no acaban en el nodo raíz S por lo que acaban siendo descartados

AMBIGÜEDAD EN EL ANÁLISIS

- Uno de los problemas más frecuentes en el análisis es que un texto puede tener múltiples interpretaciones
- ¿Cómo elegir la correcta?
- Algunos tipos de ambigüedad sintáctica son:
  - Nivel de una frase preposicional PP (¿modifica al nombre o al verbo?)
  - Coordinación y su ambigüedad estructural (and / or)
  - Ambigüedad en delimitadores

¡GRACIAS!
¿PREGUNTAS?