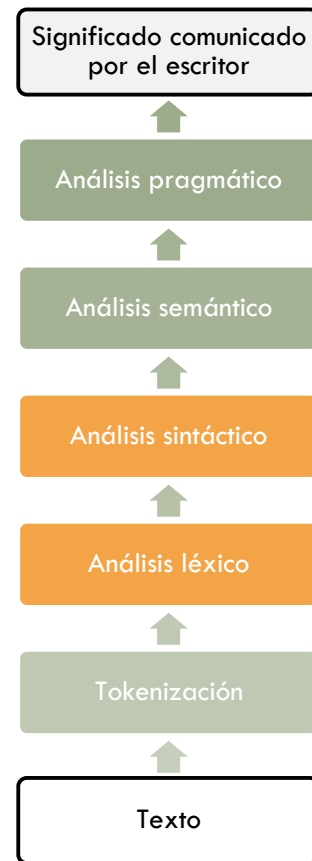


# Análisis de textos

## Análisis morfosintáctico clásico

# Análisis morfosintáctico

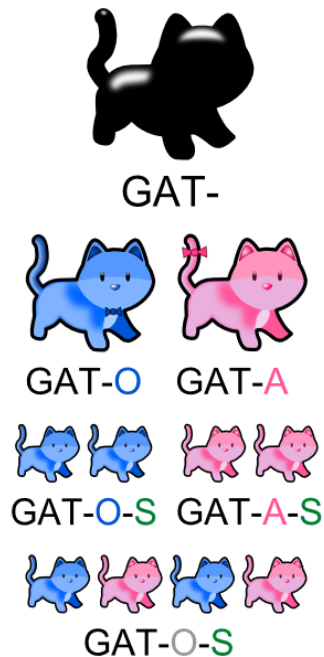
- Aunque el léxico (o morfología) y la sintaxis pueden considerarse estratos separados del lenguaje, guardan fuertes relaciones entre ellos.
- Análisis morfosintáctico: análisis léxico + análisis sintáctico.
- En este capítulo revisaremos estos dos niveles, y veremos estrategias de lingüística clásica para exponer su información.



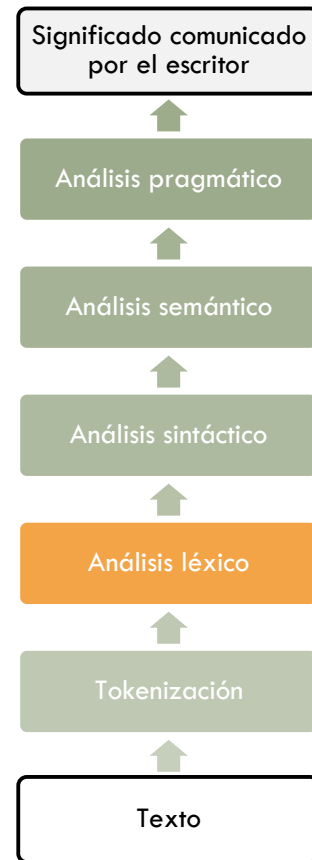
# Análisis léxico

# Análisis léxico

- Estudio de la **morfología**: la estructura interna de las palabras y su proceso de formación



- Descomposición de cada palabra en
  - **Raíz**: codifica el **significado** de la palabra
  - **Afijos** o gramemas: codifican **variantes** y **flexiones** de la palabra, como número, género, tiempo verbal, etc...
- También cálculo del **lexema** o lema: forma por defecto de la palabra (infinitivo, singular masculino, ...)



# Categorías gramaticales (Part of Speech)

- Categorías de palabras que tiene propiedades gramaticales similares
  - Palabras de la misma categoría suelen cumplir **roles similares** en la sintaxis de la frase
  - También suelen mostrar **flexiones morfológicas parecidas**
- Dependiente del idioma, aunque existen **categorías universales**

V · T · E		Lexical categories and their features	<span>hide</span>
<b>Noun</b>		Abstract / Concrete · Adjectival · Agent · Animate / Inanimate · Attributive · Common / Proper · Countable / Mass / Collective · Initial-stress-derived · Relational · Strong / Weak · Verbal / Deverbal	
<b>Verb</b>	<b>Forms</b>	Finite · Non-finite — Attributive · Converb · Gerund · Gerundive · Infinitive · Participle (adjectival · adverbial) · Supine · Verbal noun	
	<b>Types</b>	Accusative · Ambitransitive · Andative/Venitive · Anticausative · Autocausative · Auxiliary · Captative · Catenative · Compound · Copular · Defective · Denominal · Deponent · Ditransitive · Dynamic · ECM · Ergative · Frequentative · Impersonal · Inchoative · Intransitive · Irregular · Lexical · Light · Modal · Monotransitive · Negative · Performative · Phrasal · Predicative · Preterite-present · Reflexive · Regular · Separable · Stative · Stretched · Strong · Transitive · Unaccusative · Unergative · Weak	
<b>Adjective</b>		Collateral · Demonstrative · Nominalized · Possessive · Postpositive	
<b>Adverb</b>		Genitive · Conjunctive · Flat · Locative · Interrogative · Prepositional · Pronominal · Relative	
<b>Pronoun</b>		Demonstrative · Disjunctive · Distributive · Donkey · Dummy · Formal/Informal · Gender-neutral · Gender-specific · Inclusive/Exclusive · Indefinite · Intensive · Interrogative · Objective · Personal · Possessive · Prepositional · Reciprocal · Reflexive · Relative · Resumptive · Subjective · Weak	
<b>Preposition/postposition</b>		Inflected · Casally modulated · Stranded	
<b>Determiner</b>		Article · Demonstrative · Interrogative · Possessive · Quantifier	
<b>Classifier</b>		Measure word	
<b>Particle</b>		Discourse · Interrogative · Modal · Noun · Possessive	
<b>Other</b>		Copula · Converb · Expletive · Interjection (verbal) · Measure word · Preverb · Pro-form · Pro-sentence · Pro-verb · Procedure word	

Lexical categories: [https://en.wikipedia.org/wiki/Template:Lexical\\_categories](https://en.wikipedia.org/wiki/Template:Lexical_categories)

# Categorías gramaticales (Part of Speech)

## ➤ Sustantivos (o nombres)

- Palabras que refieren a un objeto o conjunto de objetos específicos

- El **gato** negro se sentó cómodamente en la **alfombra**
- **Platón** fue un **filósofo** de la antigua **Grecia**
- Compré una **bicicleta** que andaba mal

➤ Sustantivos

# Categorías gramaticales (Part of Speech)

## ➤ Verbos

- Palabras que refieren a acciones, ocurrencia o estados de existencia

- El **gato** negro se **sentó** cómodamente en la **alfombra**
- **Platón** **fue** un **filósofo** de la antigua **Grecia**
- **Compré** una **bicicleta** que **andaba** mal

➤ **Sustantivos**

➤ **Verbos**

# Categorías gramaticales (Part of Speech)

## ➤ Adjetivos

- Palabras que califican un **sustantivo**

- El **gato negro** se **sentó** cómodamente en la **alfombra**
- **Platón** fue un **filósofo** de la **antigua Grecia**
- Compré una **bicicleta** que **andaba** mal

- **Sustantivos**
- **Verbos**
- **Adjetivos**



# Categorías gramaticales (Part of Speech)

## ➤ Determinantes

- Palabras o afijos que complementan a un nombre, referenciándolo en el contexto

- El gato negro se sentó cómodamente en la alfombra
- Platón fue un filósofo de la antigua Grecia
- Compré una bicicleta que andaba mal

- Sustantivos
- Verbos
- Adjetivos
- Determinantes

# Categorías gramaticales (Part of Speech)

## ➤ Adverbios

- Palabras que modifican un verbo, adjetivo, determinante u otro adverbio

- El gato negro se sentó cómodamente en la alfombra
- Platón fue un filósofo de la antigua Grecia
- Compré una bicicleta que andaba mal

- Sustantivos
- Verbos
- Adjetivos
- Determinantes
- Adverbios

# Categorías gramaticales (Part of Speech)

## ➤ Preposiciones

- Palabras que expresan relaciones espaciales o temporales, o marcan roles semánticos

- El gato negro se sentó cómodamente en la alfombra
- Platón fue un filósofo de la antigua Grecia
- Compré una bicicleta que andaba mal

- Sustantivos
- Verbos
- Adjetivos
- Determinantes
- Adverbios
- Preposiciones

# Categorías gramaticales (Part of Speech)

## ➤ Pronombres

- Palabras que hacen referencia a un nombre en una frase
- El objeto al que refieren depende del contexto y el resto de la frase
- “Punteros a nombres”

- El gato negro se sentó cómodamente en la alfombra
- Platón fue un filósofo de la antigua Grecia
- Compré una bicicleta que andaba mal

- Sustantivos
- Verbos
- Adjetivos
- Determinantes
- Adverbios
- Preposiciones
- Pronombres

# Ejemplo de análisis léxico

## ➤ Entrada

- El gato negro se sentó cómodamente en la alfombra

## ➤ Salida

Palabra	<b>El</b>	<b>gato</b>	<b>negro</b>	<b>se</b>	<b>sentó</b>	<b>cómodamente</b>	<b>en</b>	<b>la</b>	<b>alfombra</b>	<b>.</b>
Lema	<i>el</i>	<i>gato</i>	<i>negro</i>	<i>se</i>	<i>sentar</i>	<i>cómodamente</i>	<i>en</i>	<i>el</i>	<i>alfombra</i>	<i>.</i>
Categ.	DA0MS0	NCMS000	AQ0MS0	P00CN000	VMIS3S0	RG	SPS00	DA0FS0	NCFS000	Fp

## ➤ Categoría gramatical siguiendo etiquetado estándar EAGLES

- <http://www.cs.upc.edu/~nlp/tools/parole-sp.html>

## ➤ Existen otros estándares

# Analizando el léxico de textos

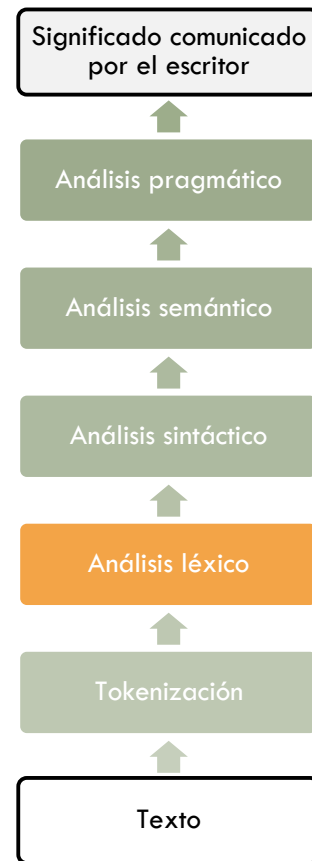
# Proceso de análisis léxico

## ➤ Entradas

- Frases, cada una de ellas dividida en tokens

## ➤ Salidas

- **Categoría gramatical** (Part of Speech) de cada token
- **Flexiones** de cada token: género, tiempo, número, etc.
- **Lexemas** de cada token
- El estudio de la morfología se basa fuertemente en la aplicación de diccionarios de palabras y reglas del idioma



# Diccionarios de morfología

- Aproximación por fuerza bruta
  - Listar en un diccionario todas las posibles formas de las palabras, anotadas según su clase morfológica y flexiones
- Ejemplo
  - gato: <SUSTANTIVO> + <MASCULINO>, <RAIZ> = gato
  - gata: <SUSTANTIVO> + <FEMENINO>, <RAIZ> = gato
  - gatos: <SUSTANTIVO> + <MASCULINO> + <PLURAL>, <RAIZ> = gato
  - gatas: <SUSTANTIVO> + <FEMENINO> + <PLURAL>, <RAIZ> = gato
  - gatos: <SUSTANTIVO> + <NEUTRO> + <PLURAL>, <RAIZ> = gato
- Desventajas
  - Obliga a repetir todas las posibles flexiones para todas las palabras
  - Poco eficiente en idiomas que admiten flexiones muy complejas pero estructuradas



# Reglas generativas de la morfología

- Pueden definirse reglas que indiquen cómo varía la forma de una palabra en función de su género, número, tiempo, etc.



GAT-



GAT-O GAT-A



GAT-O-S GAT-A-S



GAT-O-S

- Ejemplo: gato

- Gato + <MASCULINO> = Gato
- Gato + <FEMENINO> = Gata
- Gato + <MASCULINO> + <PLURAL> = Gat
- Gato + <FEMENINO> + <PLURAL> = Gatas
- Gato + <NEUTRO> + <PLURAL> = Gatos

- Se deriva la regla

- gato IS (<SUSTANTIVO>, <RAÍZ> = gat)
- <SUSTANTIVO> + <MASCULINO> = <RAÍZ> + o
- <SUSTANTIVO> + <FEMENINO> = <RAÍZ> + a
- <SUSTANTIVO> + <MASCULINO> + <PLURAL> = <RAÍZ> + o + s
- <SUSTANTIVO> + <FEMENINO> + <PLURAL> = <RAÍZ> + a + s
- <SUSTANTIVO> + <NEUTRO> + <PLURAL> = <RAÍZ> + o + s

# Reglas generativas: diversidad

➤ ¡Las reglas derivadas para gato fallan en otras palabras!

➤ Paraguas

- Paraguas + <MASCULINO> = Paraguas
- Paraguas + <FEMENINO> = <UNDEFINED>
- Paraguas + <MASCULINO> + <PLURAL> = Paraguas
- Paraguas + <FEMENINO> + <PLURAL> = <UNDEFINED>
- Paraguas + <NEUTRO> + <PLURAL> = <UNDEFINED>

➤ Nuez

- Nuez + <MASCULINO> = <UNDEFINED>
- Nuez + <FEMENINO> = Nuez
- Nuez + <MASCULINO> + <PLURAL> = <UNDEFINED>
- Nuez + <FEMENINO> + <PLURAL> = Nueces
- Nuez + <NEUTRO> + <PLURAL> = <UNDEFINED>

# Reglas generativas: subclases

- Solución: definir **subclases** morfológicas con sus reglas, y un **diccionario** de palabras asociadas

## Reglas

- **<SUSTANTIVO1>** Sustantivos regulares con ambos géneros
  - **<SUSTANTIVO1>** + **<MASCULINO>** = **<RAÍZ>** + o
  - **<SUSTANTIVO1>** + **<FEMENINO>** = **<RAÍZ>** + a
  - **<SUSTANTIVO1>** + **<MASCULINO>** + **<PLURAL>** = **<RAÍZ>** + o + s
  - **<SUSTANTIVO1>** + **<FEMENINO>** + **<PLURAL>** = **<RAÍZ>** + a + s
  - **<SUSTANTIVO1>** + **<NEUTRO>** + **<PLURAL>** = **<RAÍZ>** + o + s
- **<SUSTANTIVO2>** Sustantivos solo femeninos con plural -ces
  - **<SUSTANTIVO2>** + **<FEMENINO>** = **<RAÍZ>** + z
  - **<SUSTANTIVO2>** + **<FEMENINO>** + **<PLURAL>** = **<RAÍZ>** + ces
- **<SUSTANTIVO3>** Sustantivos solo masculinos sin flexión
  - **<SUSTANTIVO3>** + **<MASCULINO>** = **<RAÍZ>**
  - **<SUSTANTIVO3>** + **<MASCULINO>** + **<PLURAL>** = **<RAÍZ>**
- ...

## Diccionario

- gato IS (**<SUSTANTIVO1>**, **<RAIZ>**) = gat)
- perro IS (**<SUSTANTIVO1>**, **<RAIZ>**) = perr)
- nuez IS (**<SUSTANTIVO2>**, **<RAIZ>**) = nue)
- matriz IS (**<SUSTANTIVO2>**, **<RAIZ>**) = matri)
- paraguas IS (**<SUSTANTIVO3>**, **<RAIZ>**) = paraguas)
- parabrisas IS (**<SUSTANTIVO3>**, **<RAIZ>**) = parabrisas)
- ...

# Reglas generativas: análisis a través de reglas

- Para realizar el análisis léxico de una palabra
  - Se toma la palabra y se van aplicando todas las reglas posibles, a la inversa (quitando sufijos, etc...)
  - Se comprueba si alguno de los resultados obtenidos coincide con alguna entrada del diccionario y es acorde con las subcategorías morfológicas de las reglas usadas para obtenerlo
- Ejemplo: matrices
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO1} \rangle + \langle \text{MASCULINO} \rangle = \langle \text{RAÍZ} \rangle + \text{o}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO1} \rangle + \langle \text{FEMENINO} \rangle = \langle \text{RAÍZ} \rangle + \text{a}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO1} \rangle + \langle \text{MASCULINO} \rangle + \langle \text{PLURAL} \rangle = \langle \text{RAÍZ} \rangle + \text{o} + \text{s}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO1} \rangle + \langle \text{FEMENINO} \rangle + \langle \text{PLURAL} \rangle = \langle \text{RAÍZ} \rangle + \text{a} + \text{s}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO1} \rangle + \langle \text{NEUTRO} \rangle + \langle \text{PLURAL} \rangle = \langle \text{RAÍZ} \rangle + \text{o} + \text{s}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO2} \rangle + \langle \text{FEMENINO} \rangle = \langle \text{RAÍZ} \rangle + \text{z}) \rightarrow \# \text{ERROR}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO2} \rangle + \langle \text{FEMENINO} \rangle + \langle \text{PLURAL} \rangle = \langle \text{RAÍZ} \rangle + \text{ces}) \rightarrow \text{matri}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO3} \rangle + \langle \text{MASCULINO} \rangle = \langle \text{RAÍZ} \rangle) \rightarrow \text{matrices}$
  - matrices  $\rightarrow \text{INV}(\langle \text{SUSTANTIVO3} \rangle + \langle \text{MASCULINO} \rangle + \langle \text{PLURAL} \rangle = \langle \text{RAÍZ} \rangle) \rightarrow \text{matrices}$
- De las reglas que pueden aplicarse solo la primera genera una palabra de diccionario válida
  - matriz IS ( $\langle \text{SUSTANTIVO2} \rangle, \langle \text{RAIZ} \rangle = \text{matri}$ )
  - Y por la regla aplicada además sabemos que tenemos  $\langle \text{FEMENINO} \rangle + \langle \text{PLURAL} \rangle$

# Ambigüedad en el análisis léxico

- Al realizar el análisis léxico pueden encontrarse palabras de igual grafía pero clase morfológica diferente
  - Coma: <SUSTANTIVO> + <FEMENINO>, <RAIZ> = coma
  - Coma: <VERBO> + <IMPERATIVO> + <3ºPERSONA>, <RAIZ> = com
  - Coma: <VERBO> + <SUBJUNTIVO> + <1ºPERSONA>, <RAIZ> = com
  - Coma: <VERBO> + <SUBJUNTIVO> + <3ºPERSONA>, <RAIZ> = com
- Esto puede ocurrir tanto en el método de diccionario como el método basado en reglas
- Soluciones
  - Escoger la opción **más frecuente** en el idioma
  - Resolver ambigüedades en la fase de **análisis sintáctico** (más adelante)
  - Utilizar **métodos estadísticos** que comprueben palabras vecinas (más adelante)

# Extrayendo características del texto tras el análisis léxico

# Características con lemas en lugar de palabras

- Cuando lo que nos interesa es el significado de las palabras y no tanto su morfología, es preferible usar los **lemas** para extraer características
  - Ej: todas las siguientes frases indican la acción de ir al cine
    - **Fuimos** ayer al cine a ver la peli de moda
      - Lemas: <Ir ayer a el cine ver la película de moda>
    - **Fui** sola al cine y me encontré a mi prima, ¡qué coincidencia!
      - Lemas: <Ir solo a el cine y me encontrar a mi prima, ¡qué coincidencia!>
    - Tengo pensado **ir** al cine esta tarde
      - Lemas: <Tener pensar ir a el cine esta tarde>
- Por tanto, dependiendo de la aplicación es más efectivo calcular características (bag of words, n-gramas, ...) usando lemas
  - También más eficiente: #lemas < #palabras
- En algunas aplicaciones es perjudicial
  - Ej: identificación del género del escritor (fui solo VS fui sola)

# Bag-of-words y n-gramas PoS

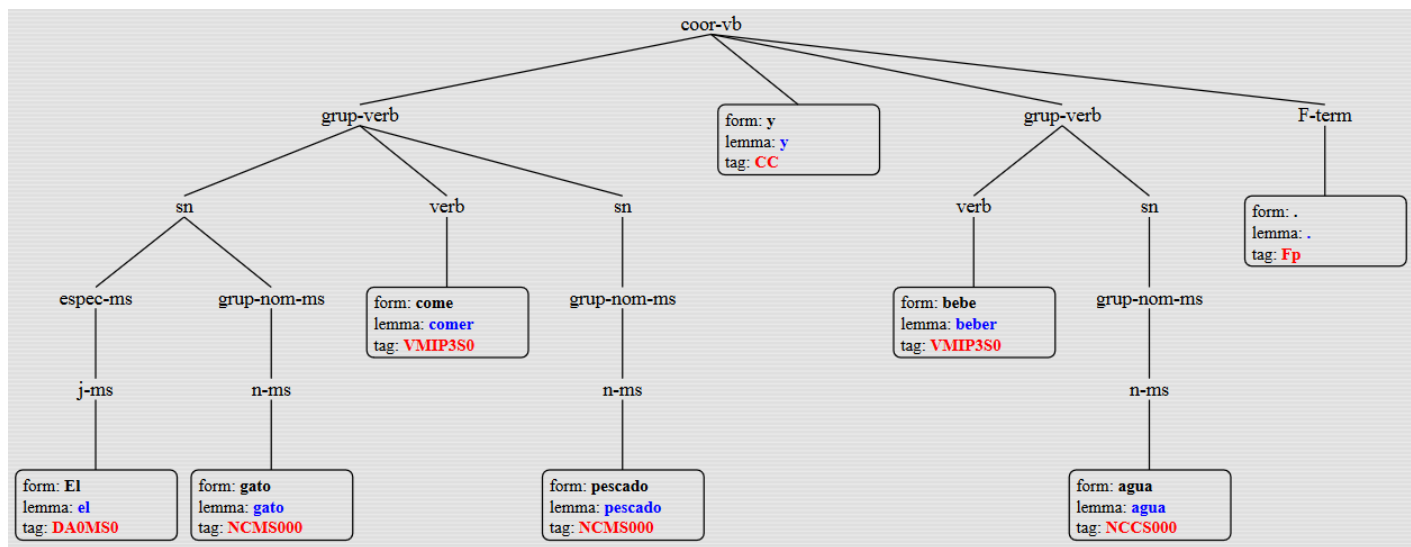
- Calcular bag-of-words o n-gramas usando las etiquetas PoS en lugar de las palabras originales
  - Proporciona información de estilo: si se usan muchos sustantivos, verbos, ...
- Puede hacerse lo mismo con otros derivados del análisis léxico, como
  - Género y/o persona de las declinaciones
  - Tiempo verbal
- Para añadir diversidad pueden considerarse tokens compuestos por palabra + PoS, y calcular bag-of-words y n-gramas sobre ellos
  - (río, SUSTANTIVO) != (río, VERBO)
- También puede ser útil filtrar todos los tokens que no sean de un PoS de interés.
  - Quedarse solo con sustantivos, verbos y adjetivos puede simplificar la identificación del tipo de documento.



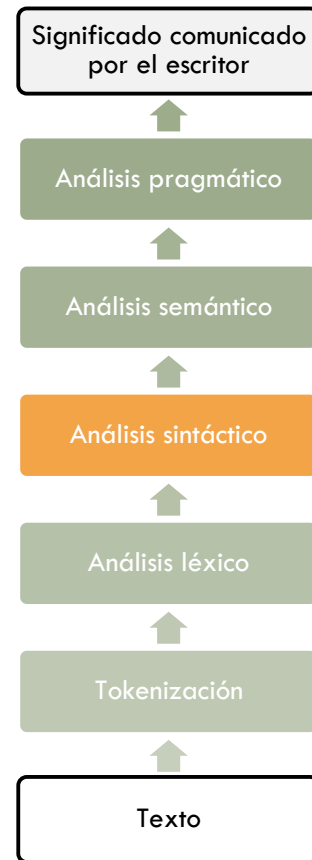
# Análisis sintáctico

# Análisis sintáctico

- Estudio de la **sintaxis**, la forma en que se combinan las palabras para formar **sintagmas** y **oraciones**
- **Árbol de parsing**: árbol de relaciones entre los componentes de una oración
- Ejemplo: “El gato come pescado y bebe agua.”



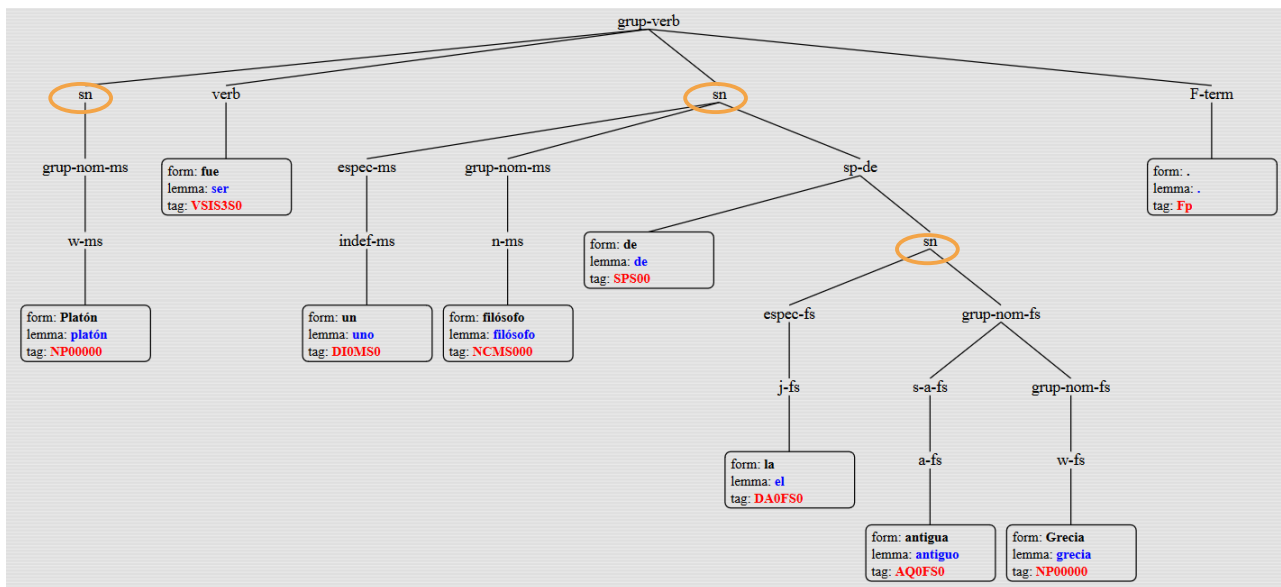
Freeling online demo: <http://nlp.lsi.upc.edu/freeling/demo/demo.php>



# Estructuras sintácticas: sintagmas

## ➤ Sintagma

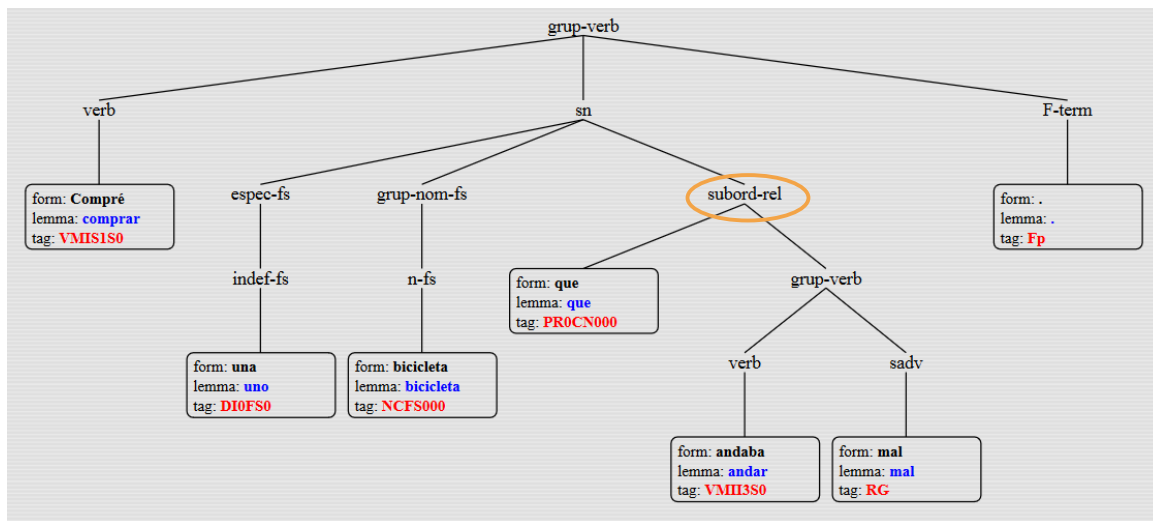
- Agrupación de palabras siguiendo la gramática del lenguaje
- Ej. Regla gramatical (Determinante + Sustantivo + Adjetivo) → sintagma (el gato negro)
- Se compone siempre de un núcleo sintáctico y otras palabras ligadas a él



# Estructuras sintácticas: oraciones subordinadas

## ➤ Oración subordinada

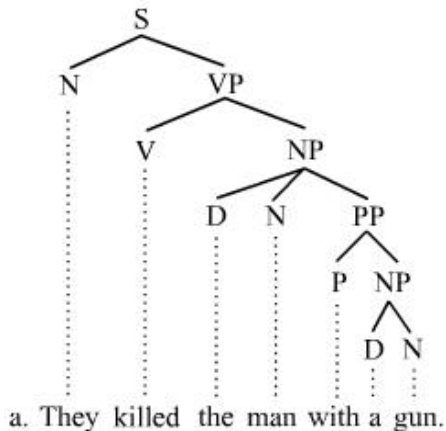
- La que depende estructuralmente del núcleo de otra oración
- No tiene autonomía sintáctica (por sí sola no es una oración bien formada)
- Ej: Compré una bicicleta que andaba mal



# Tipos de árboles de parsing

## ➤ Phrase Structure Grammar

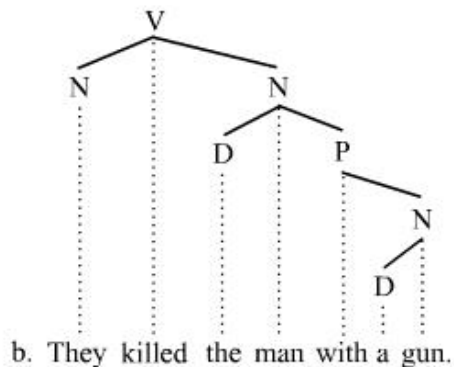
- Árbol de descomposición siguiendo una gramática
- Incluye nodos con subordinadas, sintagmas, y otros elementos que no son palabras



Phrase structure grammar

## ➤ Dependency Grammar

- Árbol que indica cómo dependen unas palabras de otras en la frase
- No incluye ningún elemento que no sean palabras
- Más simple, suele preferirse en estrategias de Machine Learning



Dependency grammar

[https://en.wikipedia.org/wiki/Constituent\\_%28linguistics%29](https://en.wikipedia.org/wiki/Constituent_%28linguistics%29)

# Analizando la sintaxis de textos

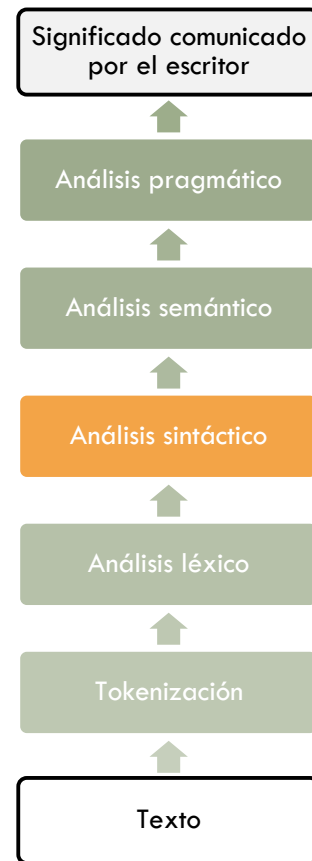
# Análisis sintáctico

## ➤ Entradas

- Una frase, con cada uno de sus tokens con Part of Speech

## ➤ Salidas

- **Árbol de parsing** expresando las relaciones sintácticas de la frase, sintagmas y oraciones subordinadas
- Para analizar la sintaxis se emplean métodos basados en **gramáticas formales**



# Gramáticas formales

- Las **gramáticas formales** establecen una serie de reglas que permiten razonar si una frase dada es gramatical (pertenece a la lengua y está bien formada)
- De entre los paradigmas de gramáticas existentes uno de los más extendidos es el de las **gramáticas generativas**
- Gramática generativa  $G = (N, T, A, R)$ 
  - **N**: conjunto de símbolos no terminales del lenguaje (sintagmas, Part of Speech, ...), incluyendo **A**
  - **T**: conjunto de símbolos terminales del lenguaje (palabras)
  - **A**: axioma, punto de partida de toda oración
  - **R**: conjunto de reglas de la gramática, que indican cómo a partir del axioma **A** se puede construir cualquiera de las oraciones válidas en la lengua

## Ejemplo

- **N**: SN | DET | N
- **T**: el | gato | barco
- **R**
  - $A \rightarrow SN$
  - $A \rightarrow N$
  - $SN \rightarrow DET\ N$
  - $N \rightarrow \text{gato} \mid \text{barco}$
  - $DET \rightarrow \text{el}$
- “el gato” → OK
- “barco” → OK
- “gato barco” → ERROR
- “gato el” → ERROR



# Clases de gramáticas: jerarquía de Chomsky

- Las gramáticas pueden clasificarse en una jerarquía en función de su poder de representación

Tipo	Gramáticas	Lenguajes	Autómatas
Tipo 0	Irrestringidas	Recursivamente enumerables	Máquinas de Turing
Tipo 1	Dependientes del contexto	Lenguajes sensibles al contexto	Autómatas lineales acotados
Tipo 2	Independientes del contexto	Independientes del contexto	Autómatas a pila
Tipo 3	Regulares o de estados finitos	Regulares	Autómatas finitos deterministas

- Cada tipo de gramática es capaz de expresar un tipo de lenguajes, que a su vez requiere de una máquina de determinada capacidad computacional para su proceso
- En general puede hacerse un análisis suficientemente preciso del lenguaje natural empleando **gramáticas independientes del contexto**

# Gramáticas independientes del contexto

- Las **gramáticas independientes del contexto** (Context Free Grammars, CFG) son una clase de gramáticas generativas en las que las **reglas** solo pueden tomar la forma
  - $A \rightarrow X_1 X_2 \dots X_n$ 
    - $A$  es un símbolo no terminal
    - $X_1 X_2 \dots X_n$  son 0 o más símbolos tanto **terminales** como **no terminales**
- Cualquier CFG puede convertirse a la **forma normal de Chomsky** (Chomsky Normal Form, CNF), que solo acepta dos tipos de reglas
  - $A \rightarrow x$
  - $A \rightarrow BC$
- La transformación a CNF puede ser difícil, por lo que se suele relajar la normalización permitiendo también reglas de la forma
  - $A \rightarrow B$

# Ejemplo de gramática CNF

- N: SNDET | SN | ADJ | DET | N
- T: el | gato | barco
- R
  - A → SNDET
  - A → SN
  - SNDET → DET SN
  - SN → N ADJ
  - SN → N
  - N → gato | barco | casa
  - DET → el | la
  - ADJ → grande | roja | pequeño

## Frases generables por la gramática

- |                  |                   |
|------------------|-------------------|
| ➤ A              | ➤ A               |
| ➤ SN             | ➤ SNDET           |
| ➤ N              | ➤ DET SN          |
| ➤ barco          | ➤ DET N ADJ       |
| ➤ A              | ➤ el casa pequeño |
| ➤ SNDET          |                   |
| ➤ DET SN         | ¡Frase gramatical |
| ➤ DET N          | pero errónea!     |
| ➤ el gato        |                   |
| ➤ A              |                   |
| ➤ SNDET          |                   |
| ➤ DET SN         |                   |
| ➤ DET N ADJ      |                   |
| ➤ la casa grande |                   |

# Gramáticas más avanzadas

- Con CFGs pueden hacerse reglas para resolver problemas de concordancia de género, número, tiempo, etc.
  - Pero el número de reglas crece demasiado
- Es mejor solución usar gramáticas más avanzadas, como las **gramáticas de unificación**
  - Cada símbolo no terminal tiene asociados una serie de atributos clave:valor
    - Ej: SN<género:masculino, número:singular>
  - Esto permite hacer reglas que mantengan la concordancia
    - SN → N ADJ si SN<género> = N<género> AND SN<género> = ADJ<género>
- De esta forma u otras similares es posible hacer gramáticas que tengan en cuenta los detalles complejos del lenguaje

# Análisis sintáctico con CFG

- Ir aplicando las reglas de la gramática de forma inversa, haciendo backtracking si no se puede llegar a una solución

➤ N: SNDET | SN | ADJ | DET | N

➤ T: el | gato | barco

➤ R

➤ A → SN

➤ A → SNDET

➤ SNDET → DET SN

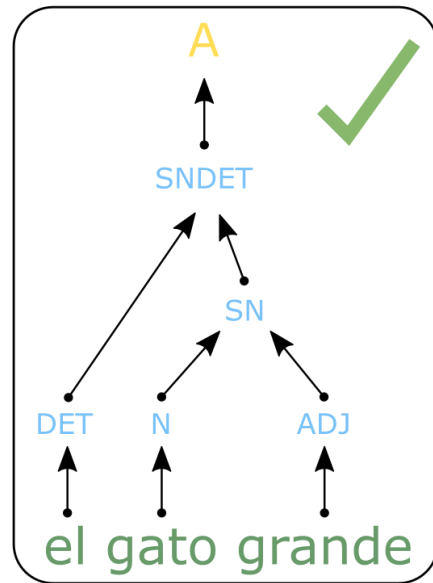
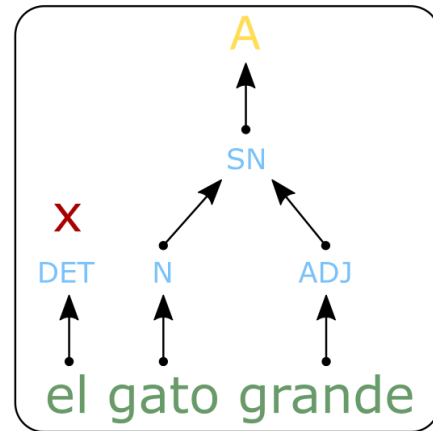
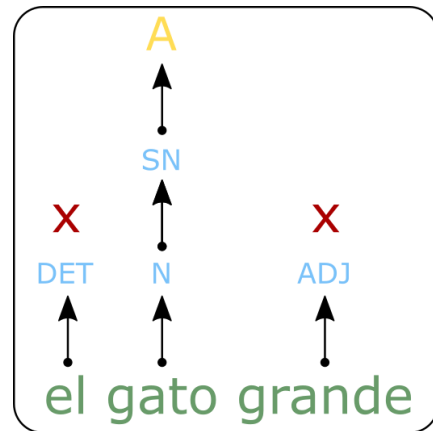
➤ SN → N

➤ SN → N ADJ

➤ N → gato | barco | casa

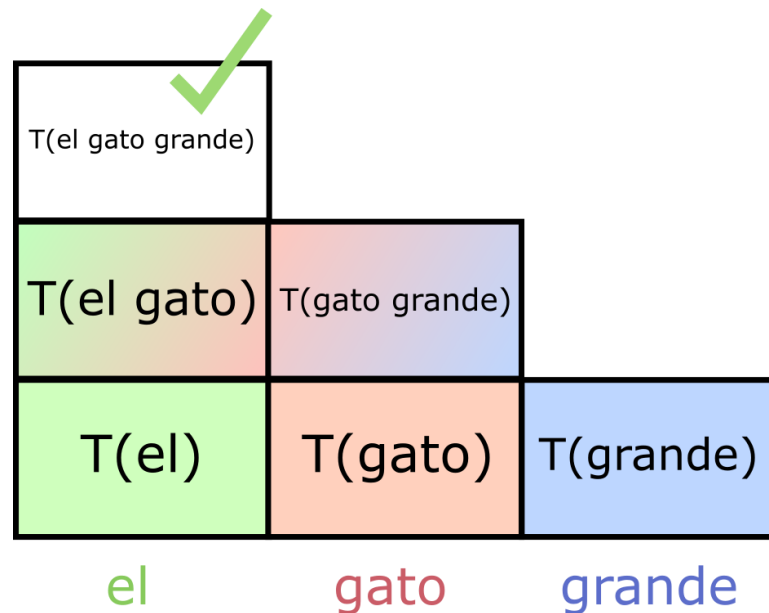
➤ DET → el | la

➤ ADJ → grande | roja | pequeño



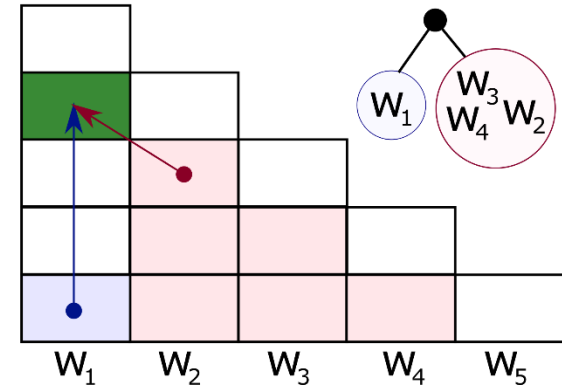
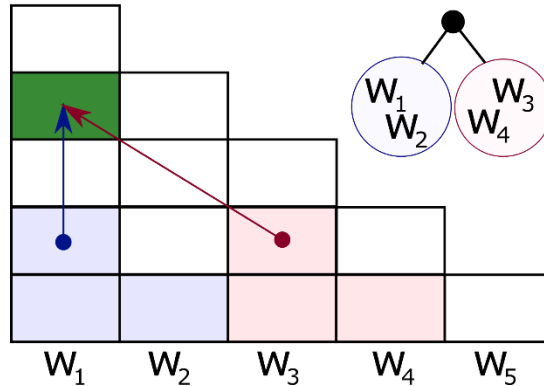
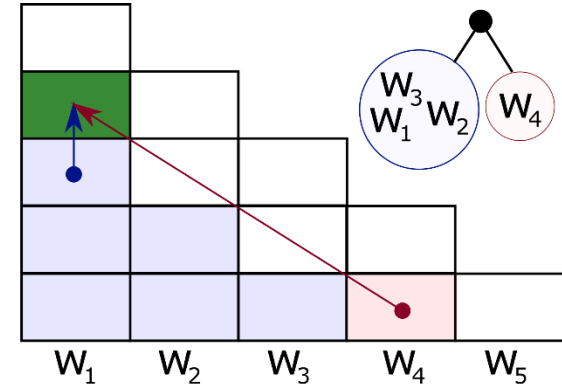
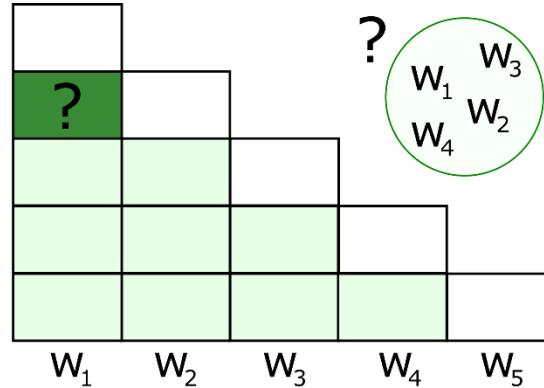
# Algoritmo de Cocke-Younger-Kasami (CYK)

- La estrategia anterior es efectiva, pero muy ineficiente
- Algoritmo de Cocke-Younger-Kasami: aceleración del análisis mediante técnicas de programación dinámica
  - Coste computacional en el caso peor  $O(n^3r)$ , con  $n$  número de palabras y  $r$  número de reglas de la CFG
  - Coste medio en la práctica  $O(n^2r)$ , para  $n=[1,30]$
- Crear una matriz diagonal inferior  $n \times n$
- Cada **columna** representa una **palabra** de la frase
- Cada **celda** representa los posibles **subárboles sintácticos** válidos que pueden construirse usando desde la palabra inferior hasta la palabra a la que llega la diagonal
- Las celdas se rellenan incrementalmente, de izquierda a derecha y de abajo a arriba
- **Inicialización**: poner en la primera fila los posibles **PoS** de cada palabra (subárboles de una sola palabra)
- **Iteración**: ir rellendo celdas mediante **combinación de subárboles**
- **Fin**: la celda superior izquierda es el árbol de toda la frase



# Algoritmo CYK: combinación de subárboles

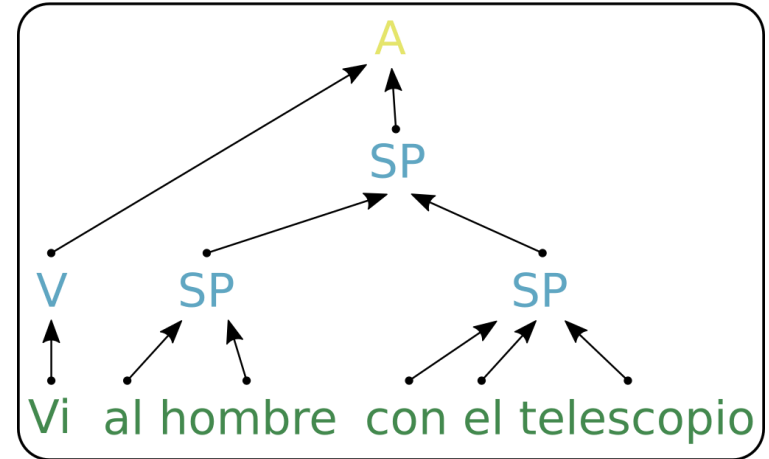
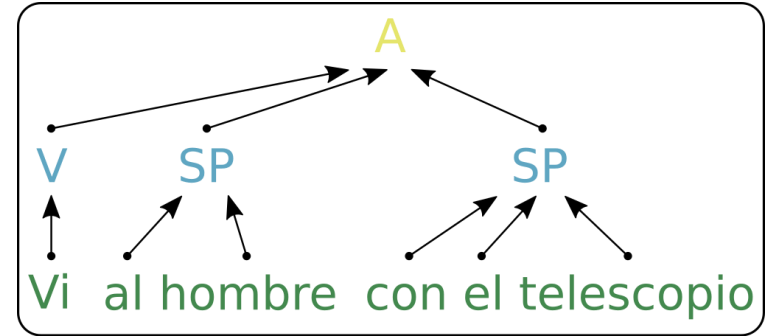
- Cada celda debe considerar todas las formas posibles de unir los subárboles creados en las **celdas de las que depende**
- Para rellenar la celda, considerar todos los posibles subárboles **sin palabras en común**, y mirar en la gramática si existe alguna regla que permita unirlos
  - Si existe, añadir el lado izquierdo de la regla a la celda
  - Si varias reglas cumplen, añadir todas
- Obs: si una celda contempla  $m$  palabras, solo pueden realizarse  $m-1$  posibles combinaciones sin palabras comunes
  - Columna (bajar) + diagonal (subir)
- Ejemplo:  
<http://lxmls.it.pt/2015/cky.html>



Ejemplo interactivo creado por Martin Lazarov

# Ambigüedad en el análisis sintáctico

- Puede existir **ambigüedad** sintáctica: varias formas diferentes de generar la frase según la gramática
  - Normalmente con significados diferentes
- En estos casos el algoritmo CYK devuelve todos los árboles sintácticos posibles
- Para desambiguar entre resultados es necesario utilizar **métodos estadísticos** que afinen los resultados obtenidos





# Gramáticas probabilísticas

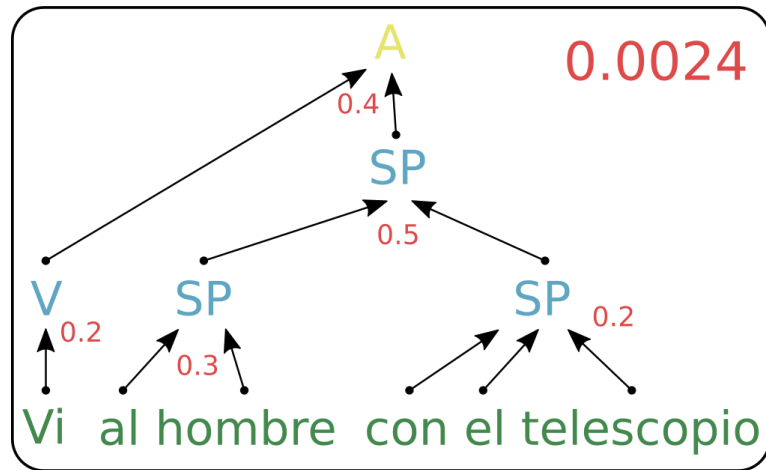
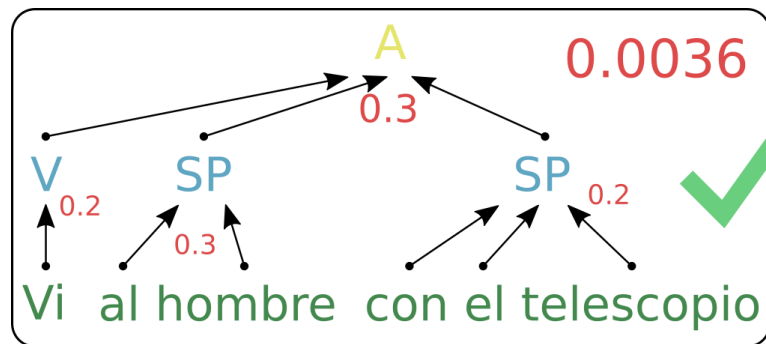
## ➤ Probabilistic Context Free Grammars (PCFG)

- Similares a CFG, pero añaden a cada regla una **probabilidad de ejecución**
- Las probabilidades de todas las reglas con mismo lado izquierdo deben sumar 1
- Puede calcularse la **probabilidad de un árbol sintáctico** completo simplemente multiplicando las probabilidades de cada una de las reglas aplicadas

- N: SNDET | SN | ADJ | DET | N
- T: el | gato | barco
- R
  - A → SNDET (0.6)
  - A → SN (0.4)
  - SNDET → DET SN (1)
  - SN → N ADJ (0.7)
  - SN → N (0.3)
  - N → gato (0.2) | barco (0.1) | casa (0.7)
  - DET → el (0.5) | la (0.5)
  - ADJ → grande (0.4) | roja (0.2) | pequeño (0.4)

# Desambiguación con PCFGs

- Para desambiguar con PCFGs se elige el árbol de **mayor probabilidad**
- Cálculo de las probabilidades: en base a **corpus anotados** sintácticamente
  - De este modo se prefieren los árboles cuya sucesión de reglas sea lo más similar posible a lo visto en el corpus de entrenamiento



# Desambiguación sintáctica avanzada

- Las PCFGs son demasiado locales para resolver todas las desambiguaciones posibles
  - Tienen en cuenta lo infrecuente de las reglas de gramática aplicadas, pero no bajo qué contexto se aplican o qué tipo de reglas se han aplicado antes
- **Solución 1:** extender la gramática para que los símbolos no terminales lleven **información sobre su origen**, y así especializar más las probabilidades
  - $A \rightarrow \text{SNDET}$  (0.6)
  - $A \rightarrow \text{SN}\langle A \rangle$  (0.4)
  - $\text{SNDET} \rightarrow \text{DET SN}\langle \text{SNDET} \rangle$  (1)
  - $\text{SN}\langle \text{SNDET} \rangle \rightarrow \text{N ADJ}$  (0.7)
  - $\text{SN}\langle \text{SNDET} \rangle \rightarrow \text{N}$  (0.3)
  - $\text{SN}\langle A \rangle \rightarrow \text{N ADJ}$  (0.6)
  - $\text{SN}\langle A \rangle \rightarrow \text{N}$  (0.4)
- **Solución 2:** usar **modelos de Machine Learning** que usen más información (más adelante)

# Extrayendo características del texto tras el análisis sintáctico

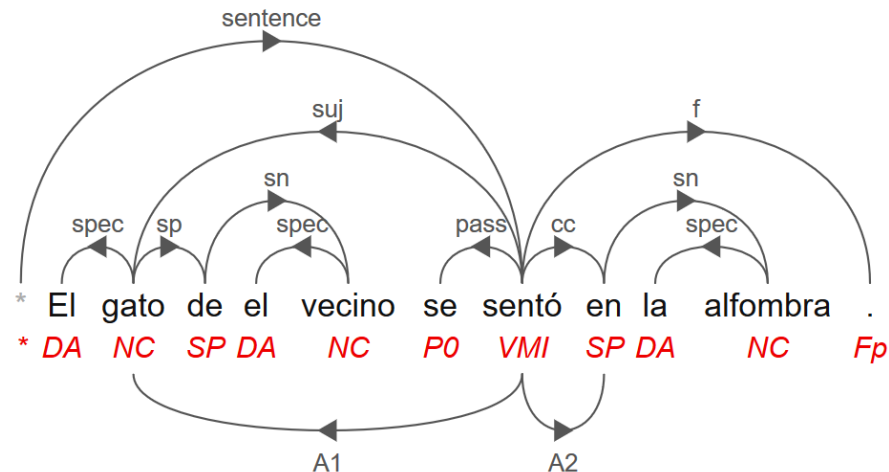
# sn-gramas (n-gramas sintácticos)

- n-gramas siguiendo árbol de **descomposición sintáctica**
  - Las palabras no se agrupan por su proximidad en la frase original, sino por cómo de cerca están en el árbol de dependencias
  - Los n-gramas se obtienen explorando el árbol en profundidad, generando n-gramas solo de arriba abajo

- Bigramas sintácticos de “El gato del vecino se sentó en la alfombra”

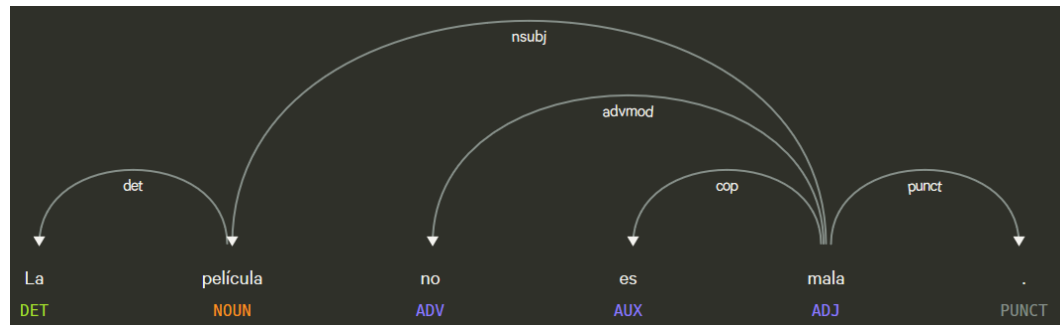
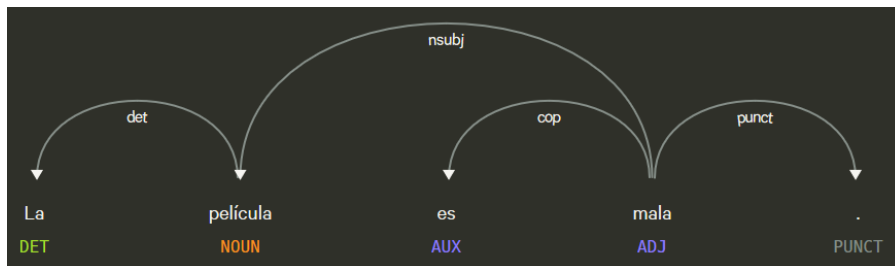
- (sentó, se) (sentó, gato) (gato, El) (gato, de) (de, vecino) (vecino, el) (sentó, en) (en, alfombra) (alfombra, la)

- Los n-gramas pueden hacerse con las palabras original, lemas, PoS, ...



# Alcance de adjetivos y adverbios

- Para análisis avanzados del texto, como el análisis de la opinión, puede ser muy útil determinar a qué palabras **influye** un adjetivo o adverbio.
- Esto puede hacerse **siguiendo el árbol** de dependencias.



# Aplicación: identificación del autor

- **Objetivo:** detectar **quién es el autor de un texto**, independientemente del dominio o tema sobre el que haya escrito
- **Características útiles** para el problema
  - Caracteres: n-gramas de caracteres, distancia de compresión (RAR)
  - Palabras: n-gramas de palabras, stop-words más frecuentes
  - Léxico: medidas de riqueza de vocabulario, uso de palabras funcionales
  - Sintaxis: n-gramas sintácticos
- **Recursos**
  - Dataset PAM: <http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-identification.html>



# Identificación del autor: ejemplo

## El último romano

Cada mañana desde hace diez o doce años, poco antes de las nueve, un hombre solitario se detiene ante la barandilla al pie del obelisco egipcio, frente al palacio de Montecitorio, en Roma, a cincuenta pasos de la entrada principal del edificio que alberga el Parlamento italiano. Es un individuo de pelo gris que ya escasea un poco, al que he visto envejecer, pues con frecuencia paso por ahí a esa hora cuando me encuentro en esta ciudad, camino del bar donde desayuno en la plaza del Panteón. Da lo mismo que sea invierno o verano, que haga sol o que llueva: apenas hay día en que no aparezca. Siempre va razonablemente vestido, con aspecto de empleado, o de funcionario. Más bien informal. Y lleva siempre una pequeña mochila, o una cartera colgada del hombro. En eso ha ido cambiando, porque ahora lo veo más con la cartera. El procedimiento es rutinario, idéntico cada día. Se detiene ante la barandilla, frente a la fachada del palacio -supongo que camino del trabajo-, saca un papel doblado que despliega con parsimonia, y con una voz sonora y educada utiliza el papel como guión o referencia de citas para el discurso que viene a continuación, diez o doce minutos de oratoria impecable, bien hilada. Un breve discurso diario, allí solo, bajo el obelisco, ante la fachada muda del Parlamento ...

## Sobre miedo, periodismo y libertad

Hace medio siglo recibí la más importante lección de periodismo de mi vida. Tenía 16 años, había decidido ser reportero, y cada tarde, al salir del colegio, empecé a frecuentar la redacción en Cartagena del diario La Verdad. Estaba al frente de esta Pepe Monerri, un clásico de las redacciones locales en los diarios de entonces, escéptico, vivo, humano. Empezó a encargarme cosas menudas, para foguearme, y un día que andaba escaso de personal me encargó que entrevistase al alcalde de la ciudad sobre un asunto de restos arqueológicos destruidos. Y cuando, abrumado por la responsabilidad, respondí que entrevistar a un político quizás era demasiado para mí, y que tenía miedo de hacerlo mal, el veterano me miró con mucha fijeza, se echó atrás en el respaldo de la silla, encendió uno de esos pitillos imprescindibles que antes fumaban los viejos periodistas, y dijo algo que no he olvidado nunca: “¿Miedo?... Mira, chaval. Cuando lleses un bloc y un bolígrafo en la mano, quien debe tener miedo es el alcalde a ti” ...

## Sobre idiotas, velos e imanes

Vaya por Dios. Compruebo que hay algunos idiotas —a ellos iba dedicado aquel artículo— a los que no gustó que dijera, hace cuatro semanas, que lo del Islam radical es la tercera guerra mundial: una guerra que a los europeos no nos resulta ajena, aunque parezca que pilla lejos, y que estamos perdiendo precisamente por idiotas; por los complejos que impiden considerar el problema y oponerle cuanto legítima y democráticamente sirve para oponerse en esta clase de cosas ...

## Libros a bordo

Hace exactamente veinte años que navego con una biblioteca a bordo. Porque una biblioteca personal, como saben ustedes, no es un lugar donde se colocan libros, sino un territorio en el que uno vive rodeado de inmediatez y de posibilidades. Hay libros que están ahí, sin leerse todavía, aguardando pacientes su momento, y otros que ya leíste y a cuyas páginas conocidas retornas en busca de memoria, de utilidad, incluso de consuelo. A medida que envejeces, el número de esa segunda clase de libros, los viejos amigos y conocidos, aumenta respecto a los que aguardan turno; aunque siempre existe la melancólica certeza de que, por mucho que vivas, nunca acabarás de leerlos todos; que la vida tiene límites, que siempre habrá libros de los que te acompañan que apenas abrirás nunca, y que un día, tanto ellos como los ya leídos caerán en manos de otros lectores: amueblarán otras vidas. Parece algo triste, pero en realidad no lo es. Porque tales son las reglas. En cierto modo, más que una vida de lecturas, una biblioteca es un proyecto de vida que nunca llegará a culminarse del todo. Eso es lo triste, y lo fascinante ...

## Baile agarrado e ira de Dios

Me ha discutido algún que otro lector la veracidad de algo que afirmé aquí hace unas semanas, cuando comparaba a nuestros hispanos curas fanáticos de antaño, o de no hace tanto, con los imanes fanáticos de hoy. En concreto, mencionaba yo el todavía reciente deseo —hace sólo setenta años— de algunos obispos españoles de meter en la cárcel a quienes bailasen agarrados, porque eso era fuente de pecado y semilla de todo mal. Y en este punto debo admitir algo: cuando lo escribí me goteaba el colmillo, clup, clup, clup, porque conozco a mis clásicos y sabía que más de uno iba a entrar a por uvas. Así que, si les parece bien, hoy vamos con ello ...

¿Mismo  
autor?



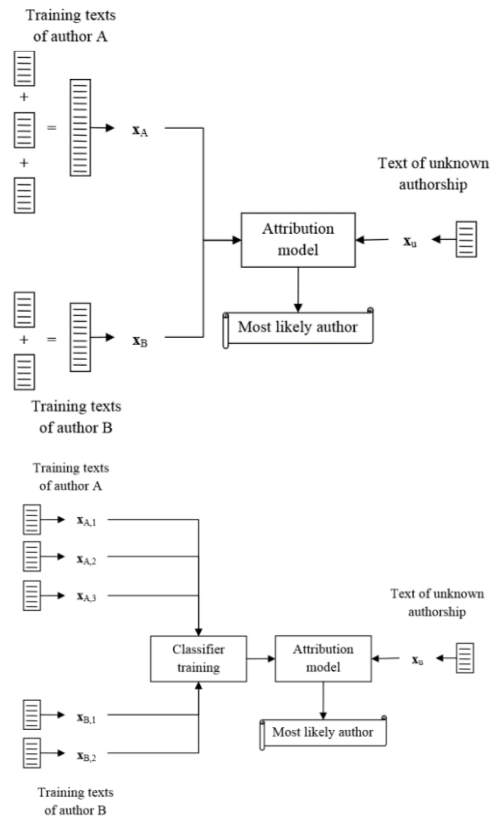
# Identificación del autor: estrategias de identificación

## ➤ Identificación basada en **perfil**

- Concatenar todos los textos de un mismo autor
- Extraer características de cada autor
- Extraer características del texto anónimo
- Atribuir al autor con menor distancia de características
  - También puede usarse distancia de compresión (RAR)

## ➤ Identificación basada en **instancias**

- Convertir cada texto en un patrón de entrenamiento, con etiqueta el autor
- Entrenar un clasificador que identifique el autor en base al texto
- Obtener el autor más probable de un texto anónimo usando el clasificador



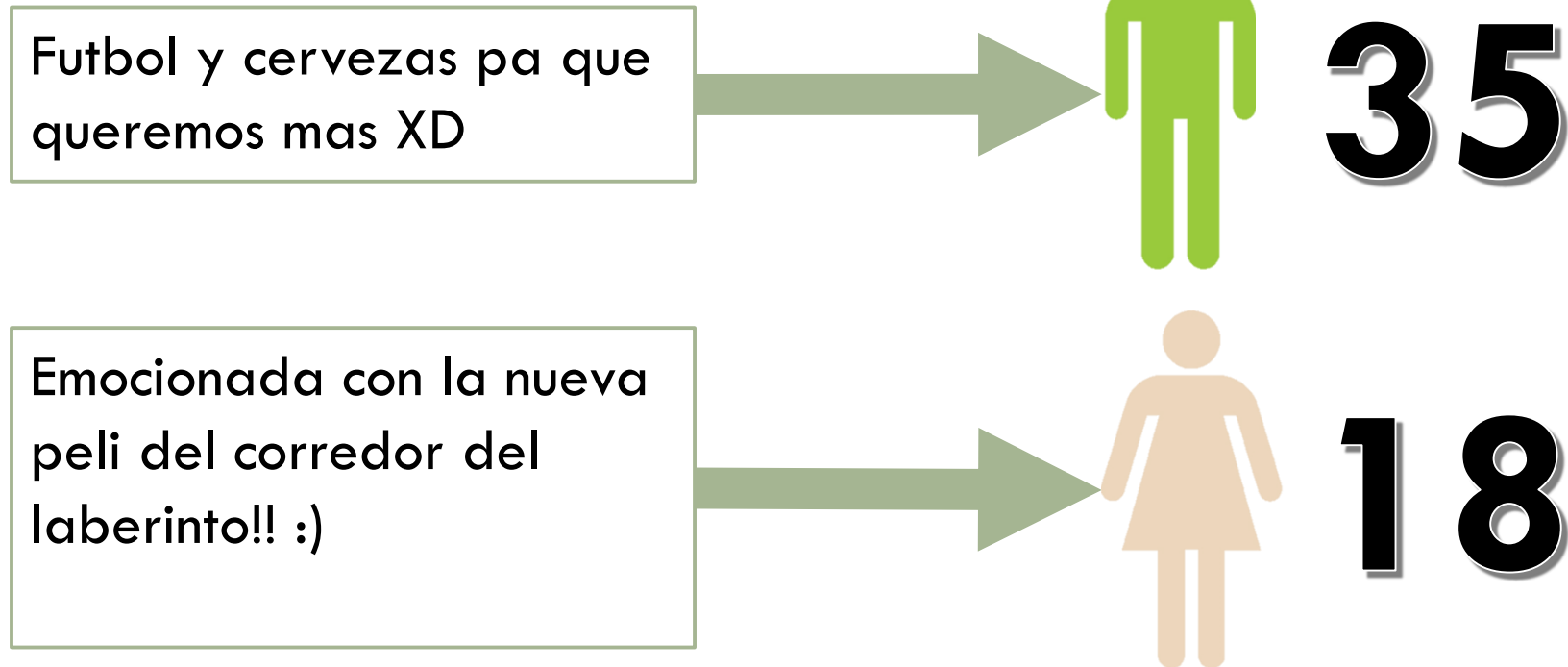
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), 538-556.

# Aplicación: perfilado de autores

- **Objetivo:** identificar género, edad, nivel educativo o rasgos psicológicos del autor: extraversión, apertura al cambio, neuroticismo, ...
- **Características útiles** para el problema
  - Caracteres: frecuencia de signos de puntuación, uso de mayúsculas, emoticonos, n-gramas de caracteres
  - Palabras: n-gramas de palabras, longitud de palabras y de frases, family tokens
  - Léxico: medidas de riqueza de vocabulario, uso de palabras funcionales, n-gramas con POS
  - Sintáctica: n-gramas sintácticos
- **Clasificadores** empleados
  - Support Vector Machines
  - Árboles de clasificación
  - Random Forest
- **Recursos**
  - Dataset PAM: <http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-web/author-profiling.html>
  - IBM Personality Insights: <https://watson-pi-demo.mybluemix.net/>



## Perfilado de autores: ejemplo



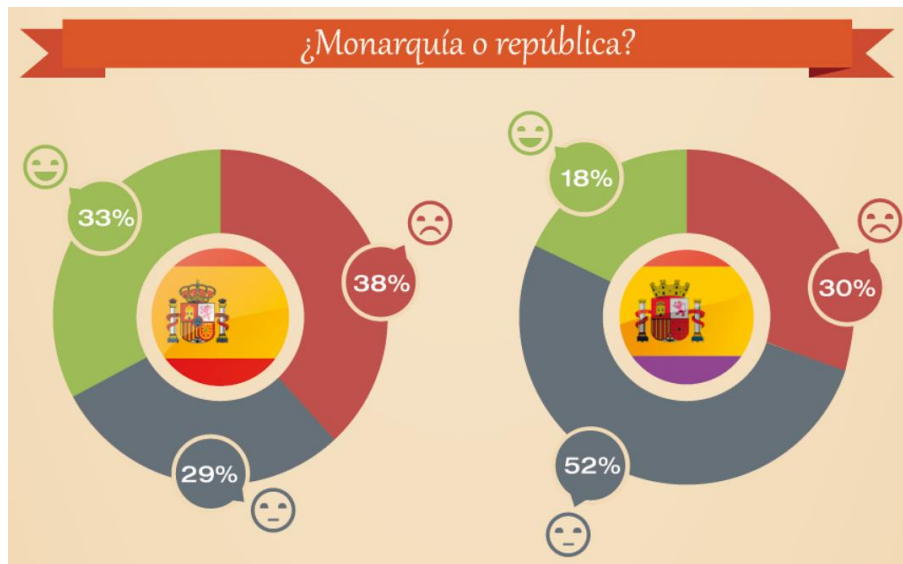
# Aplicación: análisis de opiniones

- **Objetivo:** dado un texto y un objeto determinado, detectar la opinión (positiva o negativa) del autor hacia ese objeto
  - Películas, libros, marcas, productos, personas, ...
- **Características útiles** para el problema
  - Palabras: n-gramas de palabras, diccionario de términos positivos/negativos/neutros
  - Sintáxis: árbol de parsing para detectar alcance de la negación
- **Clasificadores** empleados
  - Naive Bayes
  - Support Vector Machines




## Análisis de opiniones: ejemplo

El motor es *escaso* de potencia,  
aunque el *precio* es *muy asequible*.



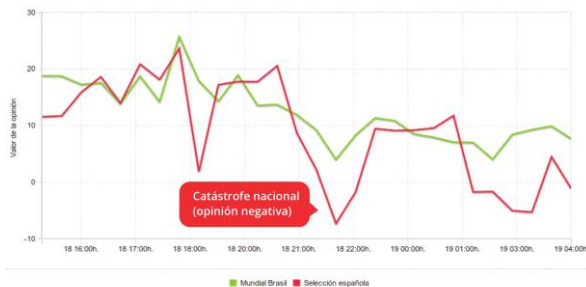
Antonio Moreno – Análisis de opinion y contenido en los medios sociales: <http://www.iic.uam.es/pdf/AnalisisDeOpinion.pdf>


## Opinión de la roja durante el Mundial

 Sentimiento hacia la selección española después del primer partido,  
**España vs Holanda**, 13 de junio.



 Sentimiento hacia la selección española durante del segundo partido,  
**España vs Chile**, 18 de junio.



 Sentimiento hacia la selección española durante del tercer partido,  
**España vs Australia**, 23 de junio.

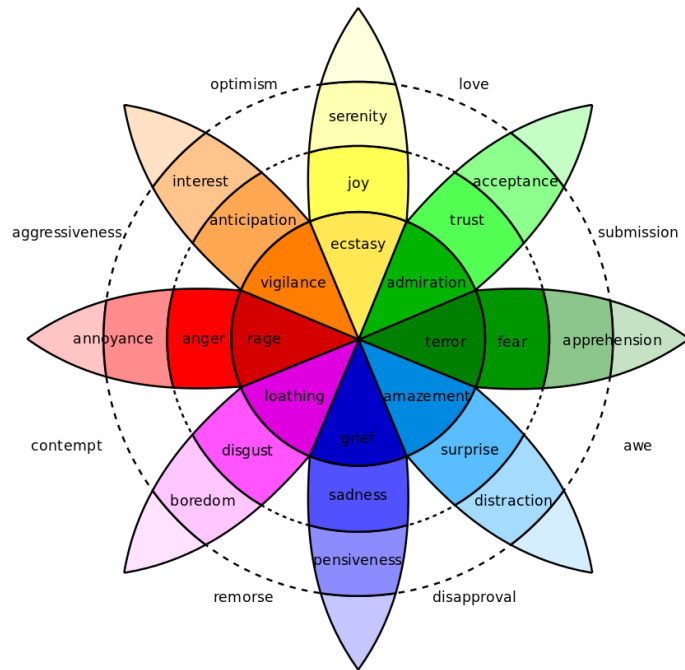


Análisis de la conversación en el Mundial 2014 con Lynguo

Análisis de la conversación en el Mundial 2014 con Lynguo: <http://www.iic.uam.es/pdf/ResumenMundial2014Lynguo.pdf>

# Aplicación: análisis de emociones

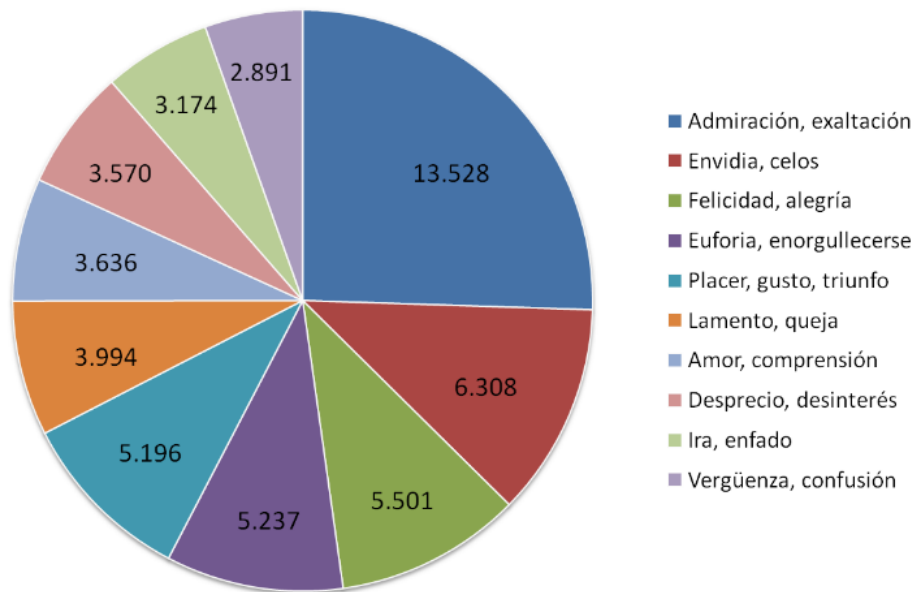
- **Objetivo:** dado un texto, detectar qué emociones o intenciones expresa el autor en el mismo
  - Ej: modelo de emociones de Plutchik
- **Características útiles** para el problema
  - Palabras: n-gramas de palabras, diccionario de términos según emociones e intensidad
  - Sintáxis: árbol de parsing para detectar alcance de la negación
- **Clasificadores** empleados
  - Naive Bayes
  - Support Vector Machines



Plutchik's wheel of emotions: [https://en.wikipedia.org/wiki/Contrasting\\_and\\_categorization\\_of\\_emotions#Plutchik.27s\\_wheel\\_of\\_emotions](https://en.wikipedia.org/wiki/Contrasting_and_categorization_of_emotions#Plutchik.27s_wheel_of_emotions)

# Análisis de emociones: ejemplo

Desglose de emociones  
#Orgullo2015



## Envidia

Siento mucha envidia mientras escucho la retransmisión del orgullo gay en Madrid, hoy era El Año! @pontechueca

Por cierto, me dais una envidia tremenda los que os habéis podido acercar a Madrid para el #Orgullo2015. El año que viene será.

## Desprecio / ira

#Orgullo2015 #OrgulloLGTB Ellos son hoy los mismos que se quejaban del corte de calles el día de la JMJ del 2011. HIPÓCRITAS DE MIERDA

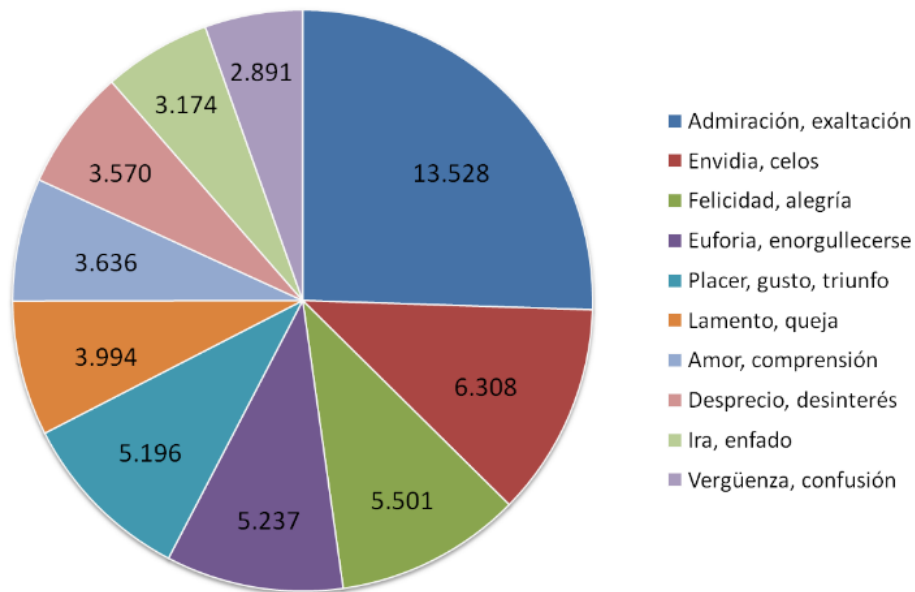
Qué puto asco me da ver la bandera del lobby #gay colgando dl Ayuntamiento de Madrid. Esa bandera NO REPRESENTA A LOS MADRILEÑOS

El #Orgullo2015, ¿más emocionante que el #24M? : [http://www.huffingtonpost.es/instituto-de-ingenieria-del-conocimiento/el-orgullo2015-mas-emocio\\_b\\_7769120.html](http://www.huffingtonpost.es/instituto-de-ingenieria-del-conocimiento/el-orgullo2015-mas-emocio_b_7769120.html)



# Análisis de emociones: ejemplo

Desglose de emociones  
#Orgullo2015



## Lamento / queja

Que los partidos políticos tengan su propio tingladito en el Orgullo Gay... da bastante polculo, con perdón.

Pues a mí el protagonismo de los políticos sean quienes sean, en el #Orgullo2015 no me gusta nada de nada..

## Asco

Los del #OrgulloLGTB han dejado el @Barriolasletras de @MADRID de suciedad y meadas, que no es para sentirse muy orgulloso...

La juventud española lo tiene claro: ¡#OrgulloGay, VERGÜENZA NACIONAL!  
#ZurulloGay <http://t.co/6HMCPPoonX>

El #Orgullo2015, ¿más emocionante que el #24M? : [http://www.huffingtonpost.es/instituto-de-ingenieria-del-conocimiento/el-orgullo2015-mas-emocio\\_b\\_7769120.html](http://www.huffingtonpost.es/instituto-de-ingenieria-del-conocimiento/el-orgullo2015-mas-emocio_b_7769120.html)



Álvaro Barbero Jiménez



@albarji



<https://github.com/albarji>



[albarji.deviantart.com](https://albarji.deviantart.com)