


Introducción al Procesamiento del Lenguaje Natural

Marta Guerrero Nieto.


Coordinadora en el Instituto de Ingeniería del conocimiento

www.iic.uam.es

A close-up photograph of several wooden blocks, similar to Scrabble tiles, scattered on a wooden surface. The blocks are light-colored wood with black letters and numbers. One block prominently shows the letter 'P' with a '4' in the bottom right corner. Another block shows the letter 'L' with a '1' in the bottom right corner. A third block shows the letter 'N' with a '1' in the bottom right corner. The text '¿Qué es el PLN?' is overlaid on the left side of the image, with '¿Qué es el' in black and 'PLN?' in purple.

¿Qué es el
PLN?

- Hace posible que las máquinas y los humanos nos comuniquemos
- Es una disciplina de la Lingüística, Inteligencia Artificial y las ciencias de la computación
- El objetivo final es comprender el lenguaje humano (semántica)
- Nació a mitad del S.XX con la traducción automática



¿Qué es la Lingüística computacional?

La Lingüística Computacional es un campo interdisciplinar que se ocupa del **desarrollo de formalismos** que describen el funcionamiento del lenguaje natural, tales que puedan ser transformados en programas ejecutables para un ordenador.

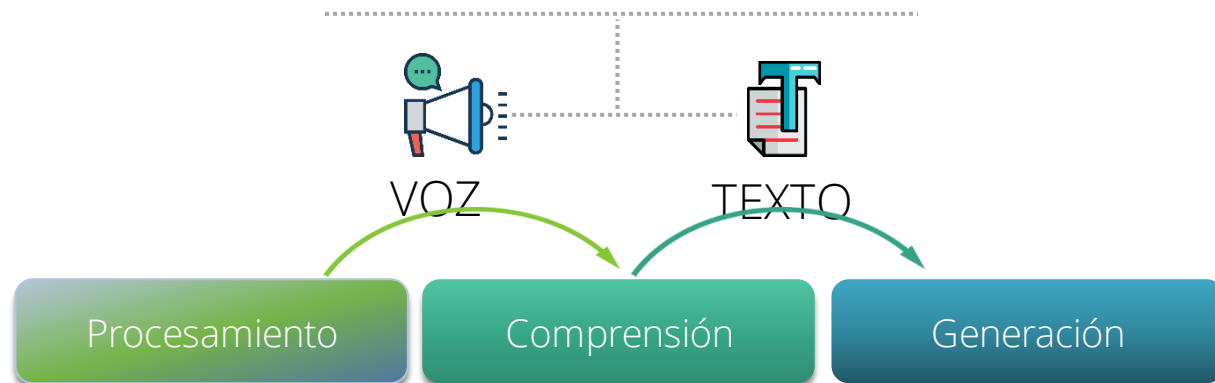
(Wikipedia)

PLN

Voz y texto



Procesamiento de Lenguaje Natural



Estructurar la información

Información estructurada

```
https://api.github.com/repos/tulios/json-viewer

1 // 20150625171327
2 // https://api.github.com/repos/tulios/json-viewer
3
4 {
5   "id": 12635853,
6   "name": "json-viewer",
7   "full_name": "tulios/json-viewer",
8   "owner": {
9     "login": "tulios",
10    "id": 33231,
11    "avatar_url": "https://avatars.githubusercontent.com/u/33231?v=3",
12    "gravatar_id": "",
13    "url": "https://api.github.com/users/tulios",
14    "html_url": "https://github.com/tulios",
15    "followers_url": "https://api.github.com/users/tulios/followers",
16    "following_url": "https://api.github.com/users/tulios/following{/other_user}",
17    "gists_url": "https://api.github.com/users/tulios/gists{/gist_id}",
18    "starred_url": "https://api.github.com/users/tulios/starred{/owner}/{repo}",
19    "subscriptions_url": "https://api.github.com/users/tulios/subscriptions",
20    "organizations_url": "https://api.github.com/users/tulios/orgs",
21    "repos_url": "https://api.github.com/users/tulios/repos",
22    "events_url": "https://api.github.com/users/tulios/events{/privacy}",
23    "received_events_url": "https://api.github.com/users/tulios/received_events",
24    "type": "User",
25    "site_admin": false
  }
```

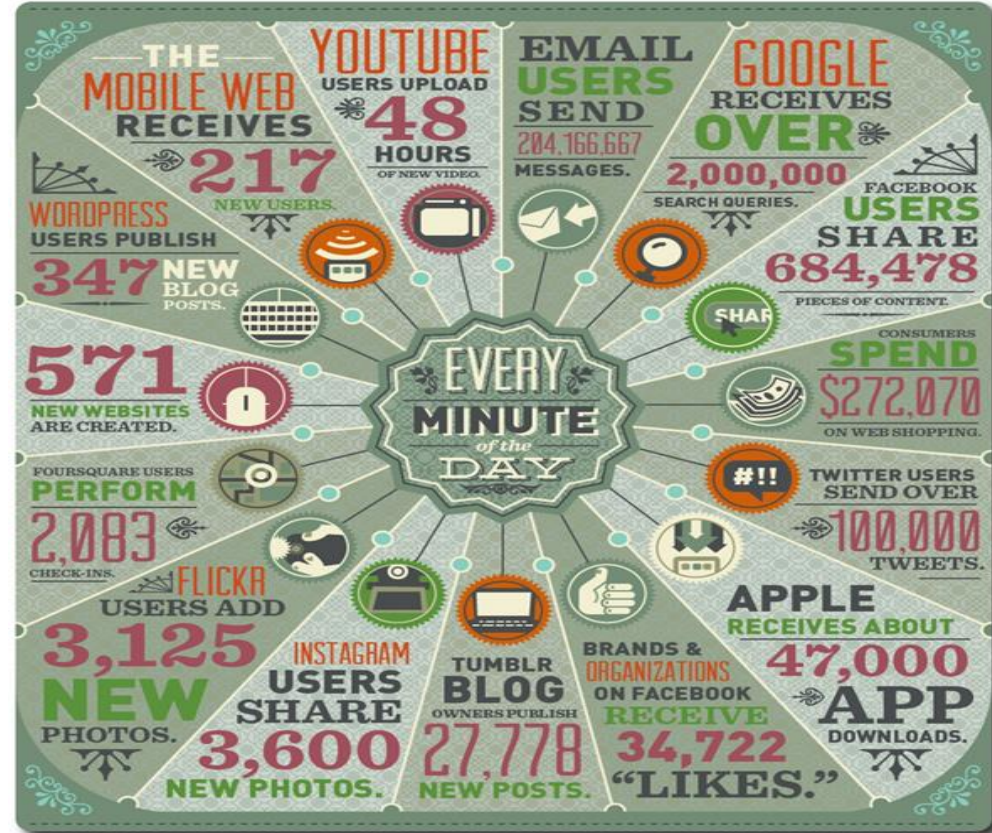
Estructurar la información

Información no estructurada

Antecedentes: apendicectomía en el año 1978, cirrosis hepática de probable origen alcohólico diagnosticada en 1989, HDA secundaria a varices esofágicas grado II en 1996 junto a hipertensión portal y ascitis. En Febrero de 1998 fue diagnosticada de hepatocarcinoma. Historia Actual: La paciente recibió trasplante hepático en octubre de 2016 que cursó sin incidencias y fue dada de alta en tratamiento con tacrolimus. 8 meses del trasplante la paciente refirió por vez primera debilidad de miembros inferiores y pérdida de sensibilidad de los mismos evolucionando en el plazo de 2 meses a una paraplejía completa que afecta a vejiga urinaria. Exploración física y pruebas complementarias: A la exploración física se observaba paraparesia con amioatrofia por desuso de EEl, hipoestesia subjetiva mayor en EID con nivel D8-D10, afectación de la sensibilidad profunda más intensa en EID sobretodo a nivel vibratoria. ROT presentes, vivos los rotulianos e hipoactivos los aquíleos. RCP extensor bilateral. En este caso también se realizaron las siguientes pruebas diagnósticas:

Información que generamos

- Se estima que el **70%-80%** de los datos que se generan son no estructurados
- Gran parte de estos son en forma de **texto libre** o lenguaje natural
- Esta información no es procesable si no es mediante técnicas especialmente diseñadas para el texto




Resumen de documentos

- ✓ Contratos
- ✓ Facturas
- ✓ Informes
- ✓ Artículos de investigación
- ✓ Noticias
- ✓ Páginas webs
- ✓ Revistas
- ✓ Emails
- ✓ Redes sociales
- ✓ Etc.

INFORMACIÓN NO
ESTRUCTURADA





¿Qué es un corpus?

¿Qué es un corpus?

Un corpus lingüístico se define por ser una colección de textos lingüísticos en formato electrónico para representar una lengua o variedad lingüística para el estudio o investigación a realizar (Sinclair, 2004).

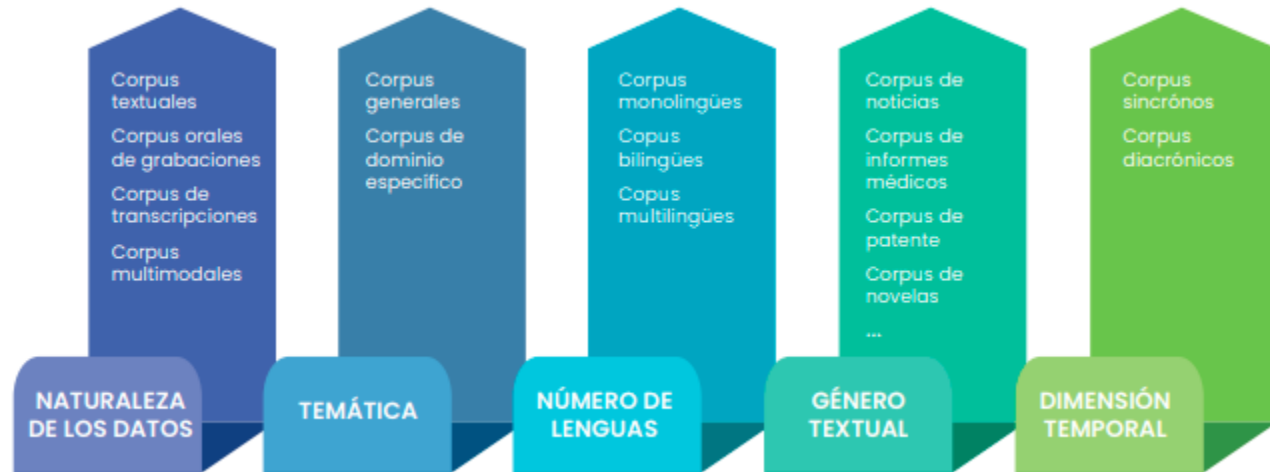
¿Qué es un corpus anotado?

Llamamos corpus anotado a aquel corpus sobre el que un **anotador humano** ha realizado una clasificación o etiquetado preciso y exhaustivo. Los criterios de anotación se establecen en la guía de anotación. Este corpus puede servir de base para entrenar **modelos de aprendizaje automático**.

Obtener un corpus anotado es costoso en cuanto a **tiempo y recursos**.

¿Qué es un corpus?

Características del corpus



Representatividad
Aspectos legales y licencias
Plataformas de anotación



Corpus en español

De donde sacamos los datos para crear un corpus anotado?

Lo habitual es que sean datos privados o internos del proyecto, pero pueden ser datos públicos.

Los más habituales son:

Prensa digital

Redes sociales (Twitter)

Wikipedia

Artículos científicos

Boletines oficiales

Páginas de derecho de la Unión europea

...

Corpus en español

- **Corpus de referencia en español (se pueden consultar, pero no descargar completos para usar en un entrenamiento)**

- • Corpus de Referencia del Español Actual (CREA):

<http://corpus.rae.es/creanet.html>

160 millones de entradas. Textos literarios y periodísticos 1975-1999

- • Corpus Diacrónico del Español (CORDE)

<http://corpus.rae.es/cordenet.html>

250 millones de entradas. Textos históricos 1250-1974

- • Corpus del Español del Siglo XXI (CORPES XXI)

<http://web.frl.es/CORPES/view/inicioExterno.view>

125 millones de entradas. Textos actuales hasta 2016

Corpus anotados en español

Corpus anotados en español

- Difícil acceso
- Gran dispersión de recursos en distintas webs
- No hay estándar
- Muchos pensados con fines de investigación

Recursos lingüísticos online:

Repositorio privado de recursos en numerosas lenguas:

<https://catalog.ldc.upenn.edu/topten>

Laboratorio de Lingüística computacional de la UAM (bajo petición)

<http://www.llf.uam.es/ESP/Recursos.html>,

CLiC- Centre de Llenguatge i de Generalitat de Catalunya (bajo petición) <http://clic.ub.edu/corpus/corpus>

Repositorio en abierto de distintos corpus anotados y no anotados:

<https://zenodo.org/search?page=1&size=20&q=spanish%20corpus>

Ejemplo de corpus (Medoccan)

Nombre: Blanca.Apellidos: Ramos Ibañez .CIPA: nhc-150679.NASS: 78 99876411 31.Domicilio: Avenida Melchor Fernández 59. 4,2.Localidad/ Provincia: Barcelona.CP: 01022. NHC: 150679.Datos asistenciales.Fecha de nacimiento: 05/03/1996.País: España.Edad: 22 Sexo: M.Fecha de Ingreso: 28/02/2016. Especialidad: Urología.Medico: Josep Rubio Palau Paseo N°Col: 08-08-25574.Motivo de ingreso: Dolor abdominal. Antecedentes: Mujer sometida a una ligadura de trompas por vía laparoscópica. Durante el mismo se detectó una tumoración de 20 mm en la cara lateral derecha de la vejiga, bien delimitada e hipoecoica. Exploración física: Se realizó un estudio ecográfico pélvico de control y una urografía intravenosa, en la cual no se detectó ninguna alteración del aparato urinario superior. Resumen de pruebas complementarias: En el cistograma de la misma se puso de manifiesto un defecto de repleción redondeado y de superficie lisa, localizado en la pared vesical derecha. Las analíticas de sangre y orina estaban dentro de los límites normales. Se le realizó una cistoscopia a la paciente, donde se objetiva la presencia de una tumoración a modo de "joroba", de superficie lisa y mucosa conservada, en cara lateral derecha de vejiga, inmediatamente por encima y delante del meato ureteral ipsilateral. Evolución y comentarios: Con el diagnóstico de presunción de leiomioma vesical se practicó resección transuretral de la tumoración. Los fragmentos resecados tenían un aspecto blanquecino, sólido y compacto, parecidos a los de un adenoma prostático, con escaso sangrado. El material obtenido de la resección transuretral estaba formado por una proliferación de células fusiformes de citoplasma alargado, al igual que el núcleo, y ligeramente eosinófilo. No se apreciaron mitosis ni atipias. El estudio inmunohistoquímico demostró la positividad para actina músculo específica (DAKO, clon HHF35) en las células proliferantes. A los tres meses de la resección transuretral se realizó cistoscopia de control, observando una placa calcárea sobreelevada sobre el área de resección previa, compatible con cistopatía incrustante que se trató mediante resección transuretral de ésta y de restos leiomiomatosos y acidificación urinaria posterior.Remitido por:Josep Rubio Palau Paseo. Av. Vall d'Hebron, 119-129 08035 Barcelona, España Email: jrubiopalau@yahoo.es

Ejemplo de corpus (Medoccan)

Nombre: Blanca. Apellidos: Ramos Ibañez .CIPA: nhc-150679.NASS: 78 99876411 31.Domicilio: **Avenida Melchor Fernández 59. 4,2**. Localidad/ Provincia: **Barcelona**.CP: **01022**. NHC: 150679.Datos asistenciales .Fecha de nacimiento: **05/03/1996**. País: España. Edad: **22** Sexo: **M**. Fecha de Ingreso: **28/02/2016**. Especialidad: Urología. Medico: **Josep Rubio Palau Paseo** N°Col: **08-08-25574**. Motivo de ingreso: Dolor abdominal. Antecedentes: Mujer sometida a una ligadura de trompas por vía laparoscópica. Durante el mismo se detectó una tumoración de 20 mm en la cara lateral derecha de la vejiga, bien delimitada e hipoeoica. Exploración física: Se realizó un estudio ecográfico pélvico de control y una urografía intravenosa, en la cual no se detectó ninguna alteración del aparato urinario superior. Resumen de pruebas complementarias: En el cistograma de la misma se puso de manifiesto un defecto de repleción redondeado y de superficie lisa, localizado en la pared vesical derecha. Las analíticas de sangre y orina estaban dentro de los límites normales. Se le realizó una cistoscopia a la paciente, donde se objetiva la presencia de una tumoración a modo de "joroba", de superficie lisa y mucosa conservada, en cara lateral derecha de vejiga, inmediatamente por encima y delante del meato ureteral ipsilateral. Evolución y comentarios: Con el diagnóstico de presunción de leiomioma vesical se practicó resección transuretral de la tumoración. Los fragmentos resecados tenían un aspecto blanquecino, sólido y compacto, parecidos a los de un adenoma prostático, con escaso sangrado. El material obtenido de la resección transuretral estaba formado por una proliferación de células fusiformes de citoplasma alargado, al igual que el núcleo, y ligeramente eosinófilo. No se apreciaron mitosis ni atipias. El estudio inmunohistoquímico demostró la positividad para actina músculo específica (DAKO, clon HHF35) en las células proliferantes. A los tres meses de la resección transuretral se realizó cistoscopia de control, observando una placa calcárea sobreelevada sobre el área de resección previa, compatible con cistopatía incrustante que se trató mediante resección transuretral de ésta y de restos leiomiomatosos y acidificación urinaria posterior.Remitido por: **Josep Rubio Palau Paseo. Av. Vall d'Hebron, 119-129 08035 Barcelona, España** Email: **jrubioalau@yahoo.es**

Ejemplo de corpus anotado (Medoccan)

T1 CORREO_ELECTRONICO 2356 2376 jrubiopalau@yahoo.es
T2 PAIS 2342 2348 España
T3 TERRITORIO 2331 2340 Barcelona
T4 TERRITORIO 2325 2330 08035
T5 CALLE 2291 2324 Paseo. Av. Vall d'Hebron, 119-129
T6 NOMBRE_PERSONAL_SANITARIO 2273 2290 Josep Rubio Palau
T7 SEXO_SUJETO_ASISTENCIA 443 448 Mujer
T8 ID_TITULACION_PERSONAL_SANITARIO 380 391 08-08-25574
T9 NOMBRE_PERSONAL_SANITARIO 342 359 Josep Rubio Palau
T10 FECHAS 298 308 28/02/2016
T11 SEXO_SUJETO_ASISTENCIA 277 278 M
T12 EDAD_SUJETO_ASISTENCIA 267 269 22
T13 PAIS 253 259 España
T14 FECHAS 235 245 05/03/1996
T15 ID_SUJETO_ASISTENCIA 185 191 150679
T16 TERRITORIO 166 171 01022
T17 TERRITORIO 151 160 Barcelona
T18 CALLE 94 127 Avenida Melchor Fernández 59. 4,2

Herramientas de anotación

- Ofrecen mayor facilidad para realizar los procesos de anotación y de armonización
- Gratuita o de pago
- Requieren una pequeña formación
- Tiene formatos de entrada y salida de datos específicas



Ejemplo de Guía de anotación del corpus

Reglas positivas

- **P1.1. Anotar el nombre propio y todos los apellidos del paciente.**

Ejemplos: Nombre: **Rafael**
Apellidos: **Calvo Martín**

- **P1.2. Anotar las abreviaturas y las iniciales, incluso aquellas que no parecen coincidir con un nombre.**

Ejemplos: Nombre: **Dña. M. Jesús**
Nombre: **Fco. José**

- **P1.3. Anotar los apodos, mote, alias, sobrenombres, hipocorísti**

Ejemplos: **[INV]** Nombre: **M. del Mar (Marita)**
[INV] Apellidos: **Vinardell**
[INV] Apellidos: **esposa del Sr. Alvarado**
[INV] **Ernesto "Che" Guevara**

- **P1.4. Anotar los títulos nobiliarios.**

Ejemplos: **[INV]** Nombre: **Duque de Alba...**

- **P1.5. Anotar el plural de los antropónimos**

Ejemplos: **[INV]** Nombre: **acudieron los Pérez para ...**

Las guías de anotación tienen reglas positivas y reglas negativas

Reglas negativas

- **N1.1. NO incluir en la etiqueta los tratamientos, por ejemplo, Sr., Sra., Dña., etc.**

Ejemplos: Nombre: **Dña. M. Jesús**

Reglas multipalabra

- **M1.1. Anotar como una sola mención los nombres y apellidos de pacientes si aparecen seguidos en el texto separados sólo por espacios o por guiones.**

Ejemplos: Nombre: **Francisco Javier**
Apellidos: **Martínez-Aguado**

- **M1.2. Anotar como una sola mención los nombres y apellidos de pacientes, aunque alguno de los dos parezca dudoso, si no se tiene información para desambiguar.**

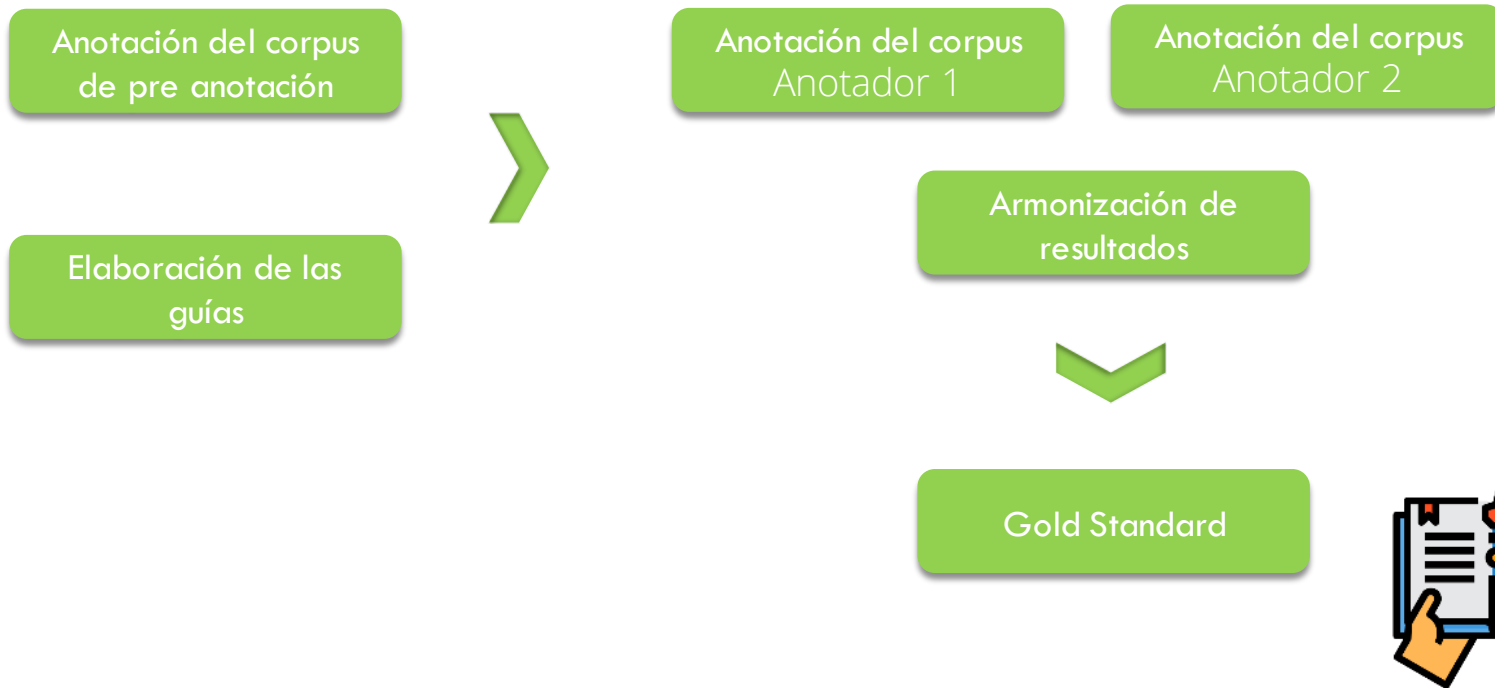
Ejemplos: **[INV]**... **francisco musulmán** de 50 años de edad ...

<http://temu.bsc.es/meddocan/wp-content/uploads/2019/02/gu%C3%ADas-de-anotaci%C3%B3n-de-informaci%C3%B3n-de-salud-protegida.pdf>

<http://temu.bsc.es/meddocan/index.php/resources/>



Metodología de anotación de un corpus



Armonización

- Proceso de resolver las discrepancias y evitar los posibles sesgos en la anotación
- gold standard

¿Qué pasa si hay mucho desacuerdo entre los dos anotadores?

Ciclo de anotación Matter

Pensado para para la creación de corpus destinados a entrenar algoritmos de lingüística computacional

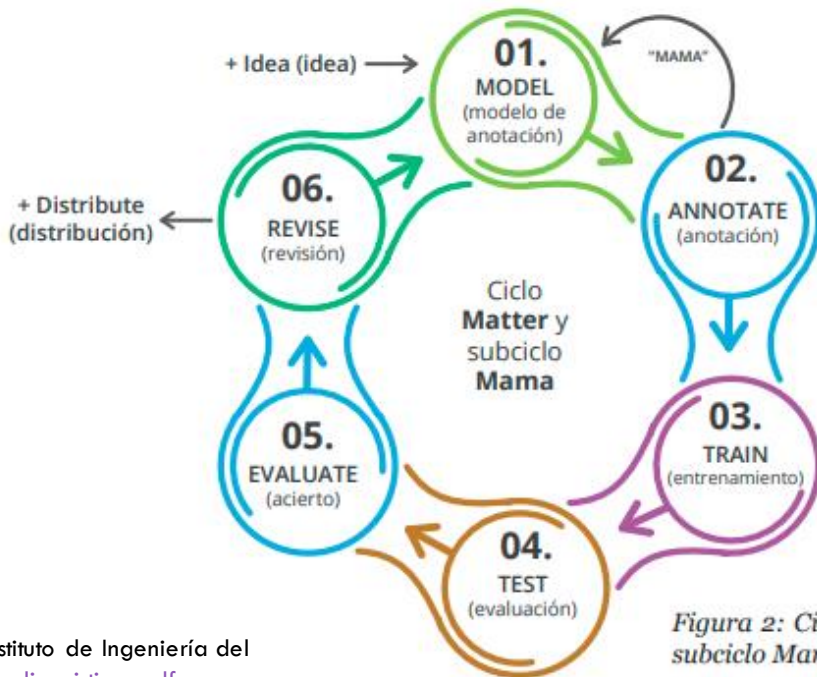


Figura 2: Ciclo Matter y subciclo Mama³

Anotación de corpus lingüísticos: metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC), <https://www.iic.uam.es/pdf/anotacion-corpus-linguisticos.pdf>

Matter

1. *Modelo (Definir la tarea a resolver, buscar bibliografía sobre la tarea y los fenómenos que se tratan en la misma, recopilar los recursos que conforman el corpus, obtener algunas métricas sobre el corpus que puedan ser de interés antes de anotar, construir el modelo, preanotar y conformar las guías de anotación)*
2. Anotación (anotación por pares, concreción en los criterios)
3. Entrenamiento (pruebas con distintos modelos)
4. Prueba (subconjunto de test)
5. Evaluación (métricas de evaluación del acierto)
6. Revisión (reanotación, revisión criterios, etc)

Consideraciones a las métricas

1. Diversidad de los datos. Un factor a tener en cuenta es la homogeneidad de los datos a anotar.
2. Similitud de las etiquetas.
3. Dificultad para caracterizar las etiquetas.
4. Diferencias entre los anotadores
5. Consistencia en las anotaciones de cada anotador

Tipos de modelos

➤ Modelos simbólicos / basados en reglas

Sistemas formales compuestos por símbolos que definen las combinaciones gramaticales de una lengua.

➤ Modelos estadísticos / probabilísticos

Sistemas basados en la frecuencia de aparición de las palabras y su probabilidad de aparecer en un contexto determinado.

Machine Learning:/aprendizaje automático

Modelos supervisados

El modelo aprende a partir de datos etiquetados.

Modelos no supervisados

El modelo infiere patrones a partir de datos no etiquetados.

Aprendizaje supervisado

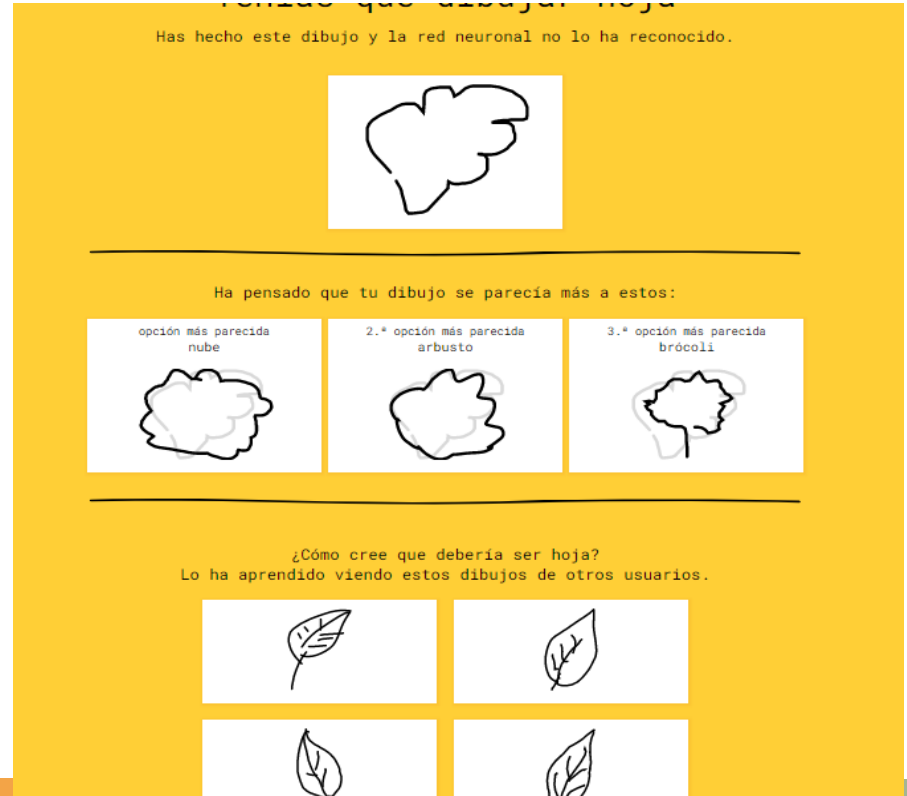
<https://quickdraw.withgoogle.com>



¿Puede una red neuronal reconocer tus dibujos?

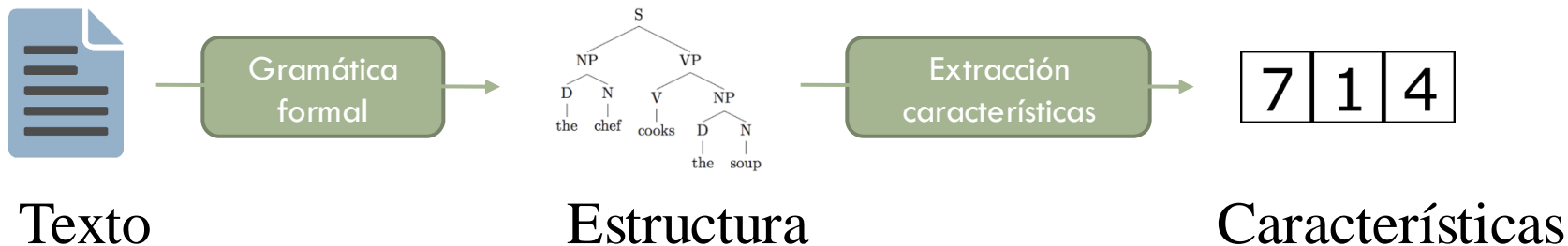
Añade tus dibujos al [conjunto de datos de dibujos más grande del mundo](#), compartido públicamente, para ayudarnos con la investigación sobre el aprendizaje automático.

¡A dibujar!



Lingüística computacional para procesamiento de datos

- El lenguaje natural es, por naturaleza, **informal**
- Para tratar adecuadamente el texto necesitamos de **formalismos** que nos permitan aproximarnos al lenguaje natural de una forma disciplinada y no ambigua
- La **lingüística computacional** se basa en definir **gramáticas formales**, **diccionarios**, **ontologías**, **corpus** y otros recursos que nos permiten realizar estos análisis, transformando el lenguaje en estructuras bien definidas que podemos procesar automáticamente



Problemas con el lenguaje I

- Identificación, información implícita, múltiples formas de decir lo mismo, la misma forma con múltiples significados (polisemia), anáforas, ambigüedad, ironía y lenguaje figurado, etc.



C A R A



La **cara** del
presentador era un
cuadro!



!Qué **cara** es esa
chaqueta!



La actriz llevaba pintada
la **cara** de un maquillaje
brillante.

Problemas con el lenguaje II

Uso de neologismos (anglicismos sobre todo), errores y vacilaciones ortográficas, uso de sublenguas (lengua oral, escrita,), léxico específico en dominios temáticos (léxico de automovil vs. Léxico de supermercados).



eslaids

slide



pinots

peanut



cartunes

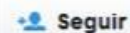
cartoon



biles

Bill

Problemas con el lenguaje III



@calisixd los abes los konstrulle el hestado, renfe solo pone trenes en la bia, rekuer2 a tu hamijo ;-))

Responder Retwittear Favorito Más

RETWEETS
6 278

FAVORITOS
3 384



17:12 - 24 de abr. de 2014

Tareas de PLN

Comprensión
NLU

Clasificación
automática

Análisis del
sentimiento

Detección de
tendencias

Traducción
automática

Extracción de información

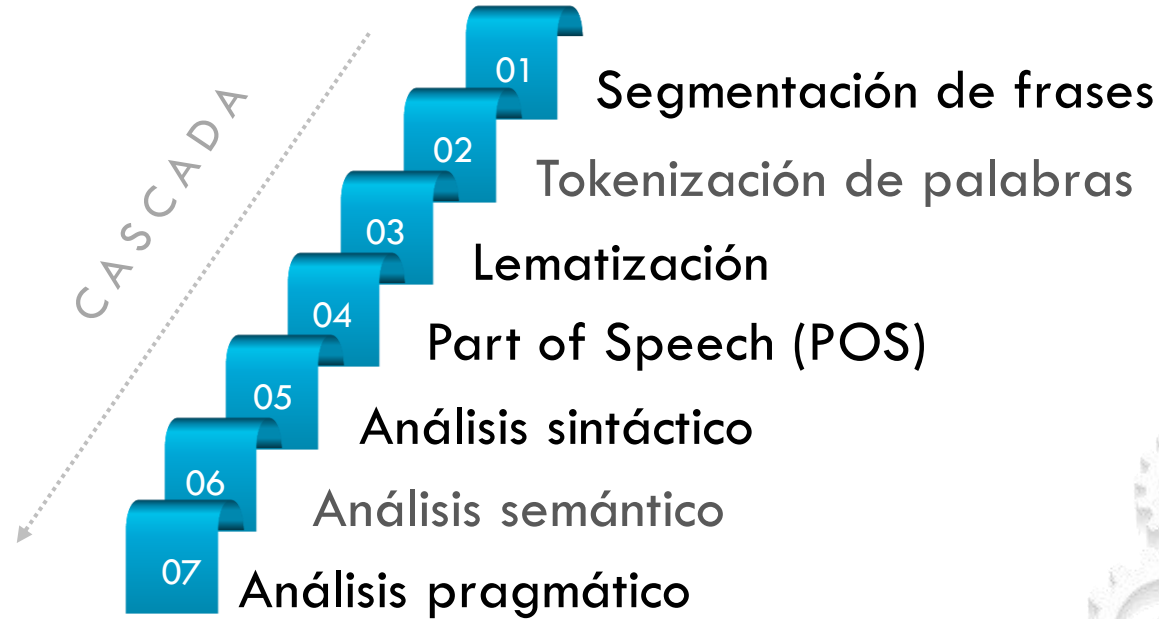
Generación
NLG

Generación automática de
texto

Chatbots

Resúmenes
automáticos

PLN CIÁSICO: CAPAS DE ANÁLISIS DEL LENGUAJE



Tokens

- Todo lenguaje escrito se conforma en base a unidades básicas
 - **Caracteres**: unidad mínima del lenguaje
 - A, i, u, k, ñ, á, ö, α, ㍻, あ, 国
 - **Fonemas**: articulación mínima de un sonido vocálico y consonántico
 - En el español: /a/, /k/, /č/ (ch), /ñ/
 - Al trabajar con texto escrito no se suelen utilizar
 - **Palabras** o **tokens**: agrupaciones de uno o varios caracteres
 - España, England, 日本
 - **Frases**: agrupaciones de palabras
- El primer paso de todo análisis del lenguaje será identificar en el texto estas unidades básicas: **tokenización**

Ejemplo de tokenización

Tokenization



Segmentación de frases / Tokenización de palabra

- Habitualmente se entiende **token** por unidad independiente separadas por espacios y por **frase** la que va separado entre puntos... pero hay abreviaturas y siglas con punto, y token normalizados compuestos de dos palabras o más.
- Problemas comunes en la tokenización:
 - **Resolución de abreviaturas** (Sr.,sra.,)
 - **Siglas** (EE.UU, RRHH,..)
 - **Términos compuestos** (Bisferol A, Buenos Aires, a cerda de,...)
 - **Expresiones coloquiales o contracciones** (padentro, pallá,...)
 - **URLs, emails, etc.**

Spacy

- Uno de las bibliotecas más importantes de PLN
- Nació en 2015 y se sigue desarrollando
- Python, Software libre
- 23 lenguas disponibles
- Tiene un curso para aprender a usarlo: <https://course.spacy.io/es/>

Ejercicio 1:

Tokenizacion con Spacy

- <https://spacy.io/usage/spacy-101>
- →python
- import spacy
- nlp = spacy.load("es_core_news_sm")
- doc = nlp(u"Desbloqueado el pacto para formar un nuevo Gobierno valenciano de coalición de PSOE, Compromís y Unides Podem")
- for token in doc:
- print(token.text)

Editable Code

spaCy v2.1.3 · Python 3 · via Binder

```
import spacy

nlp = spacy.load("es_core_news_sm")
doc = nlp(u"Desbloqueado el pacto para formar un nuevo Gobierno valenciano de coalición de PSOE, Compromís y Unides Podem")
for token in doc:
    print(token.text)
```

RUN

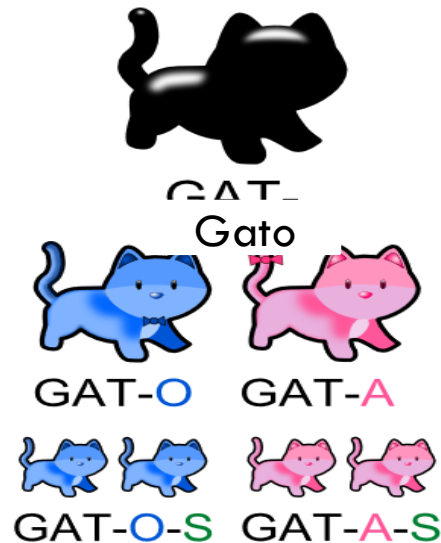
Ejercicio 1:

Tokenizacion con Spacy

- **Ejercicio 1. Probad los siguientes enunciados en Spacy (tokenización). ¿Qué problemas encontramos?**
- Disney +/ Iphone 6 /
- EE.UU / EEUU/ EE UU/ EE UU/ USA/ U.S.A
- Socio-económico/ socioeconómico/ socio_económico /email /e-mail/ e mail
- Pepitoperez.martinez@gmail.com
- 150 km/h
- www.eltiempo.es
- Gran Bretaña/ Hong Kong/ Buenos Aires
- n't (negación en inglés)
- William (Jose) Gutierrez

Lematización

- Convenio por el que se establece que todas las formas flexionadas se vinculen con una sola forma. Requiere de la información morfológica y frecuencias de apariciones para poder realizar la lematización con precisión.
- El lema es siempre el género en masculino singular o el infinitivo en el caso de los verbos.



Part of speech (POS)

- Es la asignación para cada token de su categoría gramatical correspondiente, generalmente junto con sus rasgos morfológicos.

- Dificultades

- Formato del etiquetario morfológico,
- no hay un único standard
- Entidades multipalabra, “multiwords”

ADJETIVOS			
Pos.	Atributo	Valor	Código
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participi	P

Sustantivos (y Nombres propios)

- Palabra que sirve para designar los seres vivos o las cosas materiales o mentales; gramaticalmente funciona como núcleo de un sintagma nominal, y varía en cuanto al género y al número.
- Nombres propios de persona y localizaciones

Ejemplos:

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

Verbos

- Clase de palabra con la que se expresan acciones, procesos, estados o existencia que afectan a las personas o las cosas; tiene variación de tiempo, aspecto, modo, voz, número y persona y funciona como núcleo del predicado.

Ejemplos:

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

Adjetivos

- Clase de palabra que acompaña al sustantivo para expresar una cualidad de la cosa designada por él o para determinar o limitar la extensión del mismo.

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

Determinantes

- Palabra que acompaña al sustantivo y limita o concreta su referencia, como el artículo y los adjetivos demostrativos, posesivos, indefinidos y numerales.

Ejemplos:

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

Adverbios

- Palabra invariable que modifica a un verbo, a un adjetivo, a otro adverbio o a todo un período; pueden indicar lugar, tiempo, modo, cantidad, afirmación, negación, duda y otros matices.

Ejemplos:

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

Preposiciones

- Palabra invariable que se utiliza para establecer una relación de dependencia entre dos o más palabras; la que sigue a la preposición funciona como complemento; el tipo de relación que se establece varía según la preposición.
- Las preposiciones más usuales son: a, ante, bajo, cabe, con, contra, de, desde, en, entre, hasta, hacia, para, por, según, sin, so, sobre, tras.

Ejemplos:

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

Pronombres

- Palabra que se emplea para designar una cosa sin emplear su nombre, común o propio.

Ejemplo:

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



¿Te atreves?

Señala la categoría gramatical.

*La inteligencia artificial permite
reconstruir imágenes médicas en 3D a
partir de fotografías*

Part of Speech (POS)

Sustantivos

Verbos

Adjetivos

Determinantes

Adverbios

Preposiciones

La inteligencia artificial permite reconstruir imágenes médicas en 3D a partir de fotografías

Freeling

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

▼ Sentences

Sentence 1

La	inteligencia	artificial	permite	reconstruir	imágenes	médicas	en	3D	a_partir_de	fotografías
el	inteligencia	artificial	permitir	reconstruir	imagen	médico	en	3D	a_partir_de	fotografía
DA0FS0	NCFS000	AQ0CS00	VMIP3S0	VMN0000	NCFP000	AQ0FP00	SP	Z	SP	NCFP000

► CoNLL format

1 La el DA0FS0 DA pos=determiner|type=article|gen=feminine|num=singular - - - - -
2 inteligencia inteligencia NCFS000 NC pos=noun|type=common|gen=feminine|num=singular - - - - -
3 artificial artificial AQ0CS00 AQ pos=adjective|type=qualificative|gen=common|num=singular - - - - -
4 permite permitir VMIP3S0 VMI pos=verb|type=main|mood=indicative|tense=present|person=3|num=singular - - - - - 5 reconstruir reconstruir
VMN0000 VMN pos=verb|type=main|mood=infinitive - - - - -
6 imágenes imagen NCFP000 NC pos=noun|type=common|gen=feminine|num=plural - - - - -
7 médicas médico AQ0FP00 AQ pos=adjective|type=qualificative|gen=feminine|num=plural - - - - -
8 en en SP SP pos=adposition|type=preposition - - - - -
9 3D 3D Z Z pos=number - - - - -
10 a_partir_de a_partir_de SP SP pos=adposition|type=preposition - - - - -
11 fotografías fotografía NCFP000 NC pos=noun|type=common|gen=feminine|num=plural - - - - -

Ejercicio 2: Lematización y POS en Spacy y FreeLing

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

FreeLing 4.2 - An Open-Source Suite of Language Analyzers
Enjoy the FreeLing!

Write your sentences

La inteligencia artificial permite reconstruir imágenes médicas en 3D a partir de fotografías

Select language: Auto-detect

Select output: PoS Tagging

Submit


Analysis options

- ☒ Number recognition
- ☒ Date/Time recognition
- ☒ Quantities, ratios, and percentages
- ☒ Named Entity Recognition
- ☒ Multiword detection
- ☐ Phonetic encoding
- ☒ No sense annotation
- ☐ WN sense annotation: All senses
- ☐ WN sense annotation: [UKB](#) disambiguation

- <https://spacy.io/usage/spacy-101>
- →python
- import spacy
- nlp = spacy.load("es_core_news_sm")
- doc = nlp(u"¿Por qué mueren tantos presos en EEUU?")
- for token in doc:
- print(token.text, token.lemma_, token.pos_)

Ejercicio 2: Lematización y POS

- **Probad los siguientes enunciados en Spacy y Freeling (lematización y POS)**
- 368 curados de **coronavirus** en Córdoba en un domingo sin nuevos ingresos en UCI
- Las decisiones individuales que tomemos en esta temporada de fin de año no solo afectarán a las personas más cercanas a nosotros, también afectarán a nuestras comunidades - [@Jarbas_Barbosa](#) [#COVID19](#) [#ModoSeguroDeVivir](#) [#NoBajemosLaGuardia](#)
- No le deseo mal a nadie pero si quisiera despertar un día y escuchar en las noticias q murieron d **covid 19** todos los integrantes d los 3 poderes dl estado antes q acaben con el país, económicamente y judicialmente, creo q no es delito desear q alguien se muera, sería un descanso.
- Autorizan a un millón d personas para ver el cadáver d [#Maradona](#) en un lugar cerrado. Escuelas cerradas × el [#Covid_19](#) y los chicos + brutos que ayer pero menos q mañana. El mismo día [@alferdez](#) decreta q las sesiones legislativas seguirán virtuales hasta el 21 de marzo.
- Y será que esa vacuna se la pondrá una persona que no aya presentado síntomas al **covid-19** o se la podrán poner a niños o de que edad a que edad se podrá? Me gustaría saber



Ejercicio 2: **Lematización y POS**

Freeling

- Difícil de instalar, más pesado
- Interfaz de usuario y resultados más intuitivos
- Suele desambiguar mejor el POS
- Presencia de multiwords
- La licencia v3 es libre para uso comercial

Spacy

- Ofrece resultados más simplificados

Comparar los resultados entre las herramientas ofrece dificultades porque utilizan estándar diferentes

Análisis sintáctico

✓ Dependency ✓ Parse label ✓ Part of speech ✓ Lemma Morphology

det	nsubj	amod	root	xcomp	doj	amod
La	inteligencia	artificial	permite	reconstruir	imágenes	médicas
DET	NOUN	ADJ	VERB	VERB	NOUN	ADJ
			permitir		imagen	médico

- El etiquetado sintáctico (*parsing*) es la detección de sintagmas y la asignación a cada palabra de su función oracional en relación con el resto. Dos enfoques:

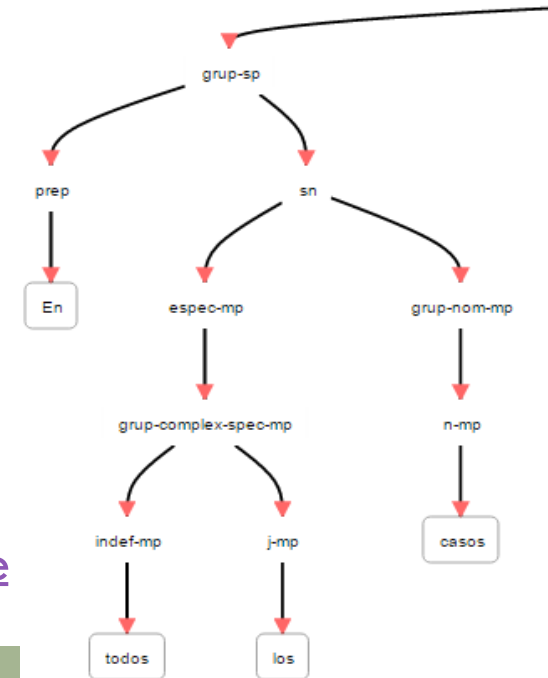
Gramática de constituyentes

Árbol de derivación sintáctica de jerarquía estructural

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

Sintaxis de dependencias

Relaciones jerárquicas léxicas entre dos palabras, no estructurales <https://cloud.google.com/natural-language>



Ejercicio 3: Sintaxis en Google Cloud

- <https://cloud.google.com/natural-language/>
Demostración de la API Natural Language

Try the API

La inteligencia artificial permite reconstruir imágenes médicas


RESET

See supported languages

Entities Sentiment **Syntax** Categories

Dependency ☒ Parse label ☒ Part of speech ☒ Lemma ☒ Morphology

det	nsubj	amod	root	xcomp	dobj
La	inteligencia	artificial	permite	reconstruir	imágenes
EI			permitir		imagen
DET	NOUN	ADJ	VERB	VERB	NOUN
gender=FEMININE number=SINGULAR proper=NOT_PROPER	gender=FEMININE number=SINGULAR proper=NOT_PROPER	gender=FEMININE number=SINGULAR proper=NOT_PROPER	aspect=IMPERFECTIVE mood=INDICATIVE number=SINGULAR person=THIRD proper=NOT_PROPER tense=PRESENT voice=ACTIVE	aspect=IMPERFECTIVE proper=NOT_PROPER voice=ACTIVE	gender=FEMININE number=PLURAL proper=NOT_PROPER



Ejercicio 3: sintaxis

Probad los siguientes enunciados en Freeling y Google Cloud (sintaxis) y comprobad las diferencias.

- El PP busca cuadrar el puzle para gobernar Madrid sin que Cs esté en la foto con Vox
- La batalla por el trono del 'streaming' continúa: ¿está Disney+ destinada a reinar entre las plataformas?
- Claves para entender el motivo de las protestas masivas en Hong Kong
- Oakland se convierte en la segunda ciudad de EEUU en despenalizar las setas alucinógenas
- ¿Por qué los coches pueden alcanzar más de 200 km/h? La UE intenta ponerles freno
- Latinoamérica: hacia la tercera ola del siglo XXI



Ejercicio 3: sintaxis

Google

- Modelo entrenado
- Interfaz muy amigable
- Servicio de análisis sintáctico con bastante precisión

Freeling

- Modelo simbólico, hecho con gramáticas
- Resulta menos intuitivo (estructura de árbol con muchos niveles artificiales)
- Resulta más sencilla la ampliación de gramáticas, ya que no necesitas corpus para entrenar

La comparación entre los dos servicios es muy compleja, no es equiparable la comparación en muchos casos.

Ejercicio 3: syntaxis

Adicionalmente se puede probar Stanza
y Spacy

<https://explosion.ai/demos/displacy>

<http://stanza.run/>

Considering using Stanza on English biomedical or clinical

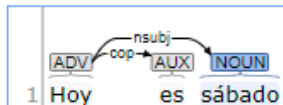
— Text to annotate —

Hoy es sábado

— Annotations —

dependency parse ✕

Universal Dependencies:



Text to parse

hoy es sábado

☒ Merge Punctuation

☒ Merge Phrases



!!Muchas gracias!!