Procesamiento de Lenguaje Natural o Minería de textos
Tema 3: Procesamiento de Lenguaje Natural.

Objetivo: El participante identificará el tipo de operaciones usadas en el pre procesamiento de textos, con la finalidad de descubrir patrones en la colección de textos y darle estructura a los mismos para su posterior tratamiento computacional.

Temario:

1. Tareas básicas: conteo de palabras, tokenización, normalización, extracción de raíces, lematización, división de oraciones
2. Etiquetado gramatical y Análisis sintáctico
3. Herramientas para usar con Python (NLTK, FreeLing, Spacy, Stanza)

Lecturas:

Applied Text Analysis with Python / by Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda : O'Reilly Media, Inc. [2018] 1 recurso en línea (xii, 334 páginas) : ilustraciones https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/

Natural language processing recipes : unlocking text data with machine learning and deep learning using Python / Akshay Kulkarni, Adarsha Shivananda -- [Berkeley, California] : Apress, [2019].-- xxv, 234 páginas : ilustraciones

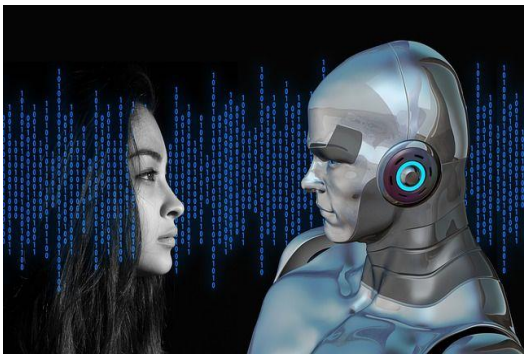
Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit 1st Edition / by Steven Bird, Ewan Klein, Edward Loper : O'Reilly Media, Inc. [2009] 1 recurso en línea (xi, 512 páginas) : ilustraciones https://itbook.store/books/9780596516499

Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. Revista de investigación de Sistemas e Informática, 6(2), 45-54.

Introducción

La mayor parte del conocimiento científico es el resultado de muchos años de investigación, con frecuencia sobre temas aparentemente no relacionados. Y lo es mucho más en las ciencias de la computación, en donde el recurso más importante que posee la raza humana es información y conocimiento.

Del procesamiento conjunto de la ciencia computacional y la lingüística aplicada, nace el **Procesamiento de Lenguaje Natural** (PLN o NLP en inglés), cuyo objetivo no es otro que el de hacer posible la compresión y procesamiento asistidos por ordenador de información expresada en lenguaje humano, o lo que es lo mismo, hacer posible la comunicación entre personas y máquinas.

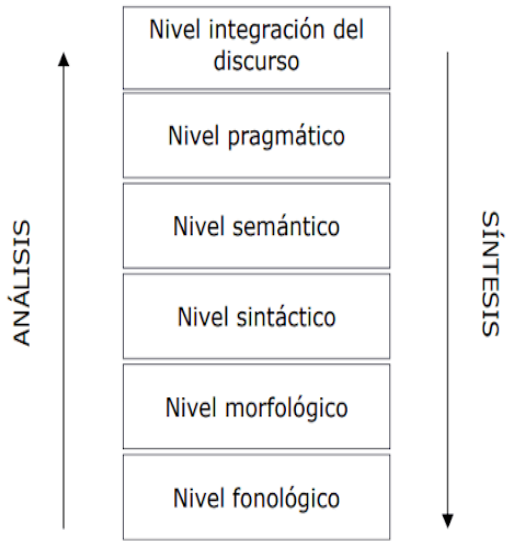


Un **lenguaje** se puede definir de diferentes formas: desde el punto de vista funcional lingüístico se define como una función que expresa pensamientos y comunicaciones entre la gente. Esta función puede realizarse mediante signos escritos (escritura) o mediante señales y vocales (voz). Desde un punto de vista formal se define como un conjunto de frases, que generalmente es infinito y se forma con combinaciones de elementos tomados de un conjunto (usualmente infinito) llamado alfabeto, respetando un conjunto de reglas de formación (sintácticas o gramaticales) y de sentido (semánticas).

Podemos distinguir entre dos clases de lenguajes: *los lenguajes naturales* (inglés, alemán, español, etc.) y *lenguajes formales* (matemático, lógico, programable etc.).

Un **lenguaje natural** es aquel que ha evolucionado con el tiempo para fines de comunicación humana, como el español o alemán[1]. Estos lenguajes continúan su evolución sin considerar la gramática, cualquier regla se desarrolla después de sucedido el hecho. En contraste, los **lenguajes formales** (aquel que el hombre ha desarrollado para expresar las situaciones que se dan en específico en cada área del conocimiento científico), están definidos por reglas preestablecidas, y por tanto se rigen con todo rigor a ellas.

Una de las tareas fundamentales de la Inteligencia Artificial (IA) es la manipulación de lenguajes naturales usando herramientas de computación, en esta, los lenguajes de programación juegan un papel importante, ya que forman el enlace necesario entre los lenguajes naturales y su manipulación por una máquina. El **PLN** consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo ésta entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje. El uso del lenguaje natural (LN) en la comunicación hombre-máquina presenta a la vez una ventaja y un obstáculo con respecto a otros medios de comunicación.



Arquitectura de un sistema de PNL

Fuente: https://itelligent.es/es/procesamiento-del-leguaje-natural-aplicaciones/

Arquitectura de un sistema de PLN

La arquitectura de un sistema de PLN se basa en una definición de Lenguaje Natural por niveles[2], los cuales son:

1. **Nivel fonológico**: trata de cómo las palabras se relacionan con los sonidos que representan
2. **Nivel morfológico**: trata de cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas

[1] [BROOKSHEAR 1993] BROOKSHEAR J. Glean.Teoría de la computación Addisson Wesley Iberoamericana Wilmington Delaware 1993.

[2] Niveles del Procesamiento del Lenguaje Natural https://youtu.be/2NfISsGZ_Rw

3. **Nivel sintáctico**: trata de cómo las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y qué sintagmas son parte de otros sintagmas
4. **Nivel semántico**: trata del significado de las palabras, y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir, de la oración aislada.
5. **Nivel pragmático**: trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones.

Esta arquitectura muestra cómo la computadora interpreta y analiza las oraciones que le sean proporcionadas:

1. El usuario le expresa a la computadora qué es lo que desea hacer.
2. La computadora analiza las oraciones proporcionadas, en el sentido morfológico y sintáctico, es decir, si las frases contienen palabras compuestas por morfemas y si la estructura de las oraciones es correcta.
3. El siguiente paso es analizar las oraciones semánticamente, es decir, saber cuál es el significado de cada oración, y asignar el significado de éstas a expresiones lógicas (cierto o falso).
4. Una vez realizado el paso anterior, ahora podemos hacer el análisis pragmático de la instrucción, es decir, una vez analizadas las oraciones, ahora se analizan todas juntas, tomando en cuenta la situación de cada oración. Una vez realizado este paso, la computadora ya sabe qué es lo que va a hacer, es decir, ya tiene la expresión final.
5. Una vez obtenida la expresión final, el siguiente paso es la ejecución de ésta, para obtener así el resultado y poder proporcionárselo al usuario.

Problemas del procesamiento de lenguaje natural

- Ambigüedad

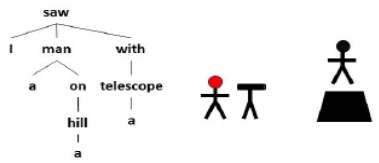
Uno de los grandes problemas del PLN se produce cuando una expresión en LN posee más de una interpretación, es decir, cuando en el lenguaje de destino se le pueden asignar dos o más expresiones distintas. Este problema de la ambigüedad se presenta en todos los niveles del lenguaje, sin excepción. Ejemplo:

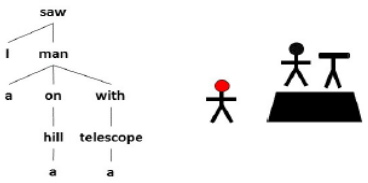
> “Hay alguien en la puerta, que te quiere hablar”
> “Hay alguien, en la puerta que te quiere hablar”

No está claro, si el predicado “te quiere hablar” se adjudica a “alguien” o a “la puerta”, sabemos que las puertas no hablan, por tanto, deducimos que es a alguien. Pero esto no lo puede deducir la máquina, a no ser que esté enterada de lo que hacen o no hacen las puertas. En apariencia este problema es demasiado sencillo, pero en realidad, es uno de los más complicados y que más complicaciones ha dado para que el PLN pueda desarrollarse por completo, ya que, al presentarse en todos los niveles del lenguaje, se tienen que desarrollar programas (lenguaje formal) para solucionarlos en cada caso.

Ambigüedad sintáctica

- Ambigüedad





I saw a man with a telescope

Ambigüedad léxico-semántica

- ¿Se quedará a dormir?
- Sí.
- Quizá debería saber que la casa está encantada.
- Ah, pues dígale que a mí también me hace ilusión quedarme.

El capitán dijo: "¡Bajen las velas!" Y los de arriba se quedaron sin luz.

- ¡Qué fresca está la mañana!
- Normal, es de hoy.

- ¿Por qué los de Lepe tiran a sus hijos a un pozo?
- Porque saben que en el fondo son buenos.

- ¿Qué pasa si un elefante se queda de pie encima de una pata?
- a- Que se cae.
- b- Que el pato se queda viudo.
- c- Que aplasta a su domador.

Ambigüedad fonológica / morfosintáctica

Dígame su nombre.
- Peter O´ Brian
- Decídase por favor.

- ¡Acusado! ¡Hable ahora o calle para siempre!
- Elijo calle.

Mi marido se ha ido de ca[S]a
1. de casa
2. de caza

Plata no es. Oro tampoco. ¿Qué es?

¿Qué esconde?
¿Que es conde?

Ambigüedad pragmática

- ¡Camarero! ¿Se puede saber qué está haciendo esta mosca en mi sopa? - Mmm, yo creo que está nadando a braza, señor.”	• ¿Cómo estás? • Han perdido los Pumas
	• Golpeó el armario con un palo y lo rompió.

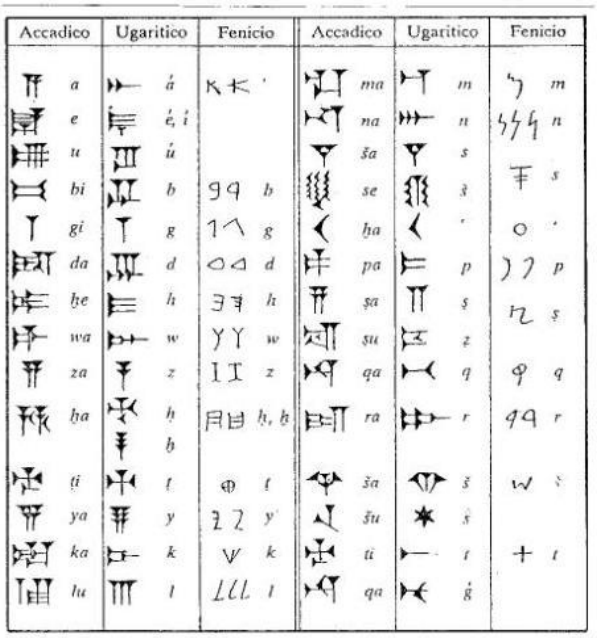
MARÍA SIMARRO VÁZQUEZ, Humor verbal basado en la ambigüedad léxica y competencia léxico-semántica, Pragmalingüística 25 (2017): 618-636

- Multiplicidad de variantes

Existen alrededor de 7.097 idiomas en el mundo, según la revista “Ethnologue”[3]. Con diferentes palabras, estructuras sintácticas, reglas morfológicas, sistemas fonéticos y escrituras. El intercambio – traducción entre unas y otras no es obvio[4].

¿Sabías que en México hay 68 lenguas indígenas, además del español?[5]

Diferentes Alfabetos



Escritura no alfabética

[3] https://www.europapress.es/sociedad/noticia-idiomas-cifras-cuantas-lenguas-hay-mundo-20190221115202.html
[4] https://wals.info/
[5] https://www.gob.mx/cultura/es/articulos/lenguas-indigenas?idiom=es

的 不 用 一 新 时 吗 所 被 说 男 事
人 在 之 二 市 也 家 后 只 着 女 得
上 于 要 三 与 还 可 分 都 位 几 真
中 前 好 四 内 出 件 种 做 把 各 对
下 者 了 五 本 去 里 将 已 吧 谁 看
大 会 年 六 地 到 最 很 长 难 找 见
小 号 月 十 这 他 回 而 行 来 子 加
是 我 日 个 此 性 万 数 等 站 字 更
没 和 为 次 建 就 能 天 再 每 那 多
有 你 名 元 全 部 爱 无 以 起 哪 少

- Evolución y cambio
  - ✓ Hay palabras que se incorporan a la lengua: tuit, celular, selfie.
  - ✓ Hay palabras que desaparecen: doncel, jumento, vuestra merced, vosotros
  - ✓ Otras cambian de sentido: hasta, celular, ratón
  - ✓ Algunas estructuras sintácticas cambian: SOV (latín) -> SVO (español)
  - ✓ Algunos fonemas (sonidos) desaparecen y aparecen otros nuevos: caballo [kabalo] > [kabaio] > [kabaʒo] / [kavaʃo]
  - ✓ Muchas lenguas desaparecen. Otras se mezclan (pidgin y criollos). Otras resucitan (hebreo), y de otras solo quedan testimonios escritos (etrusco).

- Oscuridad, slang, etc.

Amor, hagamos cuentas.
A mi edad
no es posible
engañar o engañarnos.
Fui ladrón de caminos,
tal vez,
no me arrepiento.
Un minuto profundo,
una magnolia rota
por mis dientes
y la luz de la luna
celestina.

En ocasiones los humanos, no son claros ni precisos…

https://www.poemas-del-alma.com/pablo-neruda-oda-al-amor.htm

La **lingüística computacional** es una disciplina híbrida que combina el estudio de las lenguas y la computación. Busca formular modelos que ayuden a diseñar una inteligencia lingüística. Para ello, “hay que trasladar a la máquina todos los procesos cognitivos que los seres humanos llevamos a cabo para procesar

el lenguaje", hay que transformar el lenguaje natural en un lenguaje formal, como es el de las matemáticas y la programación[6].

Tareas básicas para el procesamiento de textos

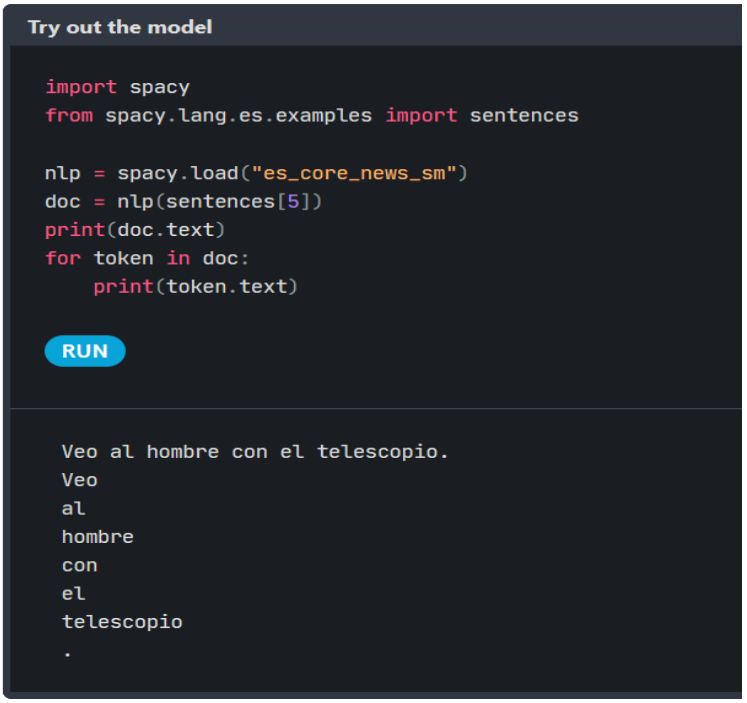
Cuando se enfrenta al texto con la idea de descubrir conocimiento, se encuentra con el problema de la falta de estructura de este. Esta falta de estructura es solo aparente, porque, realmente, el texto presenta una estructura demasiado compleja y difícil de tratar computacionalmente. Dependiendo del tipo de operaciones usadas en este preprocesamiento de datos, será el tipo de patrones a descubrir en esta colección.

Prerrequisitos para analizar un texto:

- Ser capaz de dividir el texto por frases.
- Ser capaz de encontrar las palabras.
- DEFINICIÓN DE PALABRA: Todo aquello que se encuentra entre dos espacios en blanco o espacio en blanco y signo de puntuación.

Tokenizador

- El tokenizador simplemente separa texto en una lista de tokens usando algunos caracteres como referencia para dividir.
- Los tokens son generalmente palabras y símbolos de puntuación.
- Este proceso toma en cuenta que las palabras pueden estar interrumpidas por un final de línea, están pegadas a signos de puntuación, no siempre están separadas por espacios y no siempre los espacios en blanco separan las palabras.
- La tokenización no implica ningún nivel de análisis y se realiza muy rápido.



https://spacy.io/models/es

Análisis de textos

Niveles

- Fonética-fonología (sonidos) – Corresponde al análisis de habla

[6] Leer https://elpais.com/retina/2019/01/15/tendencias/1547545169_410011.html

- Morfología (clases de palabras y segmentación)
- Sintaxis (oraciones, sintagmas y orden de las palabras)
- Semántica (significados)
- Pragmática (interacciones, uso y contexto)
- Discurso (expresiones correferenciales, estructura retórica)

Morfología

- ¿Cuáles son las clases de palabras y por qué importa saberlo?
- ¿Qué partes tienen las palabras? ¿Cómo se segmentan?

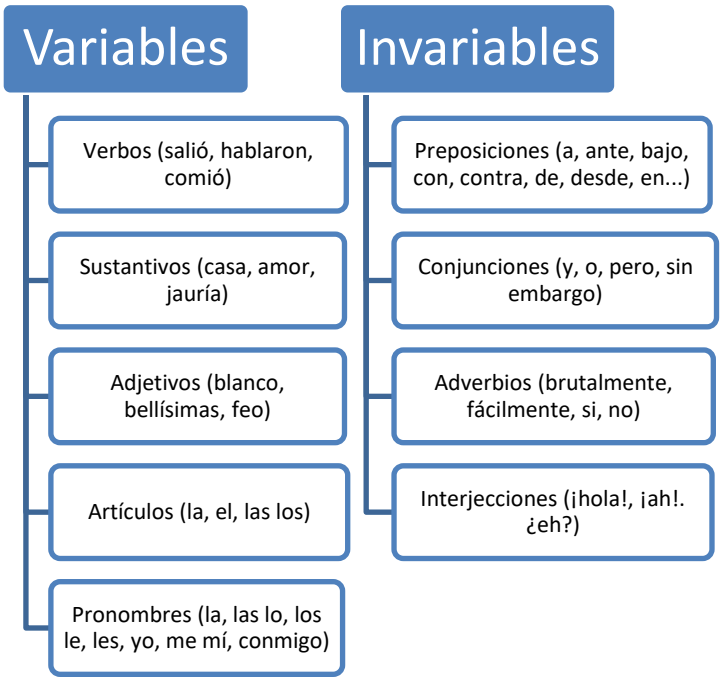
¿Cuáles son las palabras?

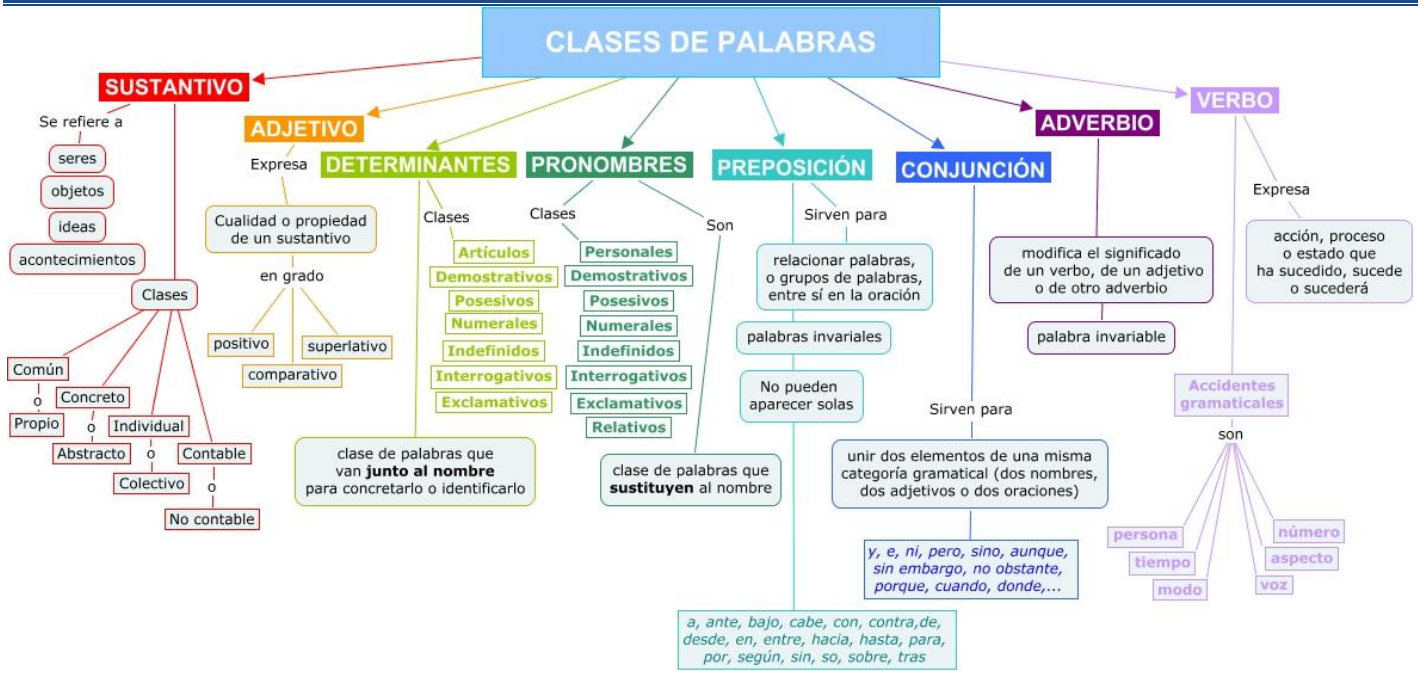
Pago por $475 347.50 M.N. (cuatrocientos setenta y cinco mil trescientos cuarenta y siete pesos 50/100 M.N.) del Restaurant Pomme de terre D'Opera por haberlo traspasado a Mary-Carmen da Cunha en San Luis Potosí.

7/02/2017

Clases de palabras

Part-of-speech (POS, etiquetado de parte del discurso), clases morfológicas, categorías gramaticales





Etiquetado Penn TreeBank

Penn TreeBank Tags está basado en el corpus Brown, pionero en etiquetado POS para inglés.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Fuente: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Etiquetas EAGLES

Las etiquetas EAGLES codifican todas las características morfológicas existentes para la mayoría de los idiomas europeos, incluido el español. Estas etiquetas consisten en un conjunto de caracteres de longitud variable donde cada uno corresponde a una característica morfológica.

https://www.cs.upc.edu/~nlp/tools/parole-sp.html

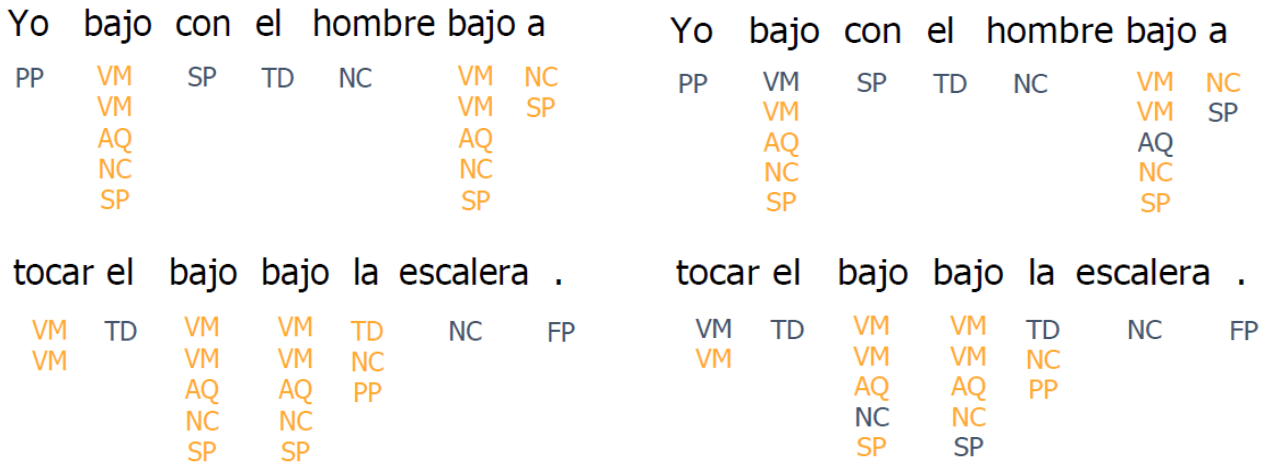
1. Adjetivos
2. Adverbios
3. Artículos
4. Determinantes
5. Nombres
6. Verbos
7. Pronombres
8. Conjunciones
9. Numerales
10. Interjecciones
11. Abreviaturas
12. Preposiciones
13. Signos de Puntuación

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	0
6	Género Semántico	-	0
7	Grado	Apreciativo	A

Forma	Lema	Etiqueta
chico	chico	NCMS000
chicos	chico	NCMP000
chica	chica	NCFS000
chicas	chica	NCFP000
oyente	oyente	NCCS000
oyentes	oyente	NCCP000
cortapapeles	cortapapeles	NCMN000
tesis	tesis	NCFN000
Antonio	antonio	NP00000

Problema del POS Tagging:

- Las palabras, tomadas en forma aislada, son ambiguas respecto a su categoría.
- Pero… La categoría de la mayoría de las palabras no es ambigua dentro de un contexto.



✓ *Solución:* Desambiguador Morfosintáctico (Pos tagger)

Desambiguador morfosintáctico es una aplicación informática en la red que realiza un análisis morfológico de palabras: lematiza cualquier palabra al identificar su forma canónica, categoría gramatical y la flexión o

derivación que la produce, y obtiene las formas correspondientes a partir de una forma canónica y de la flexión o derivación solicitada. El objetivo de un desambiguador (tagger) es el de asignar a cada palabra la categoría más "apropiada", dentro de un contexto. Basado en reglas, Estadísticos y Basados en transformaciones.

Herramientas en línea:

- ✓ Linguakit: https://linguakit.com/es/etiquetador-morfosintactico
- ✓ Stanford Parser: http://nlp.stanford.edu:8080/parser/index.jsp
- ✓ Desambiguador morfosintáctico: http://protos.dis.ulpgc.es/investigacion/desambigua/morfosintactico.htm
- ✓ Stilus: https://www.mystilus.com/herramientas/analizador-morfosintactico

Etiquetador gramatical (POS)

- El etiquetado de parte del discurso (POS, Part-of-speech) es el proceso de marcar una palabra en un texto con una parte particular del discurso, en función de su definición y contexto.
- Requieren un corpus marcado manualmente.
- Es una forma de análisis morfo-sintáctico.

https://spacy.io/models/es

Comparación de algunas herramientas[7]:

Herramienta	Código	Segmentación y tokenización	Etiquetado POS	Lematización
NLTK	Nativo Python	Si	Eagles	No
Freeling	API	Si	Eagles	Si
Pattern.es	Nativo Python	Si	Penn TreeBank	Si
Spacy	Nativo Python	Si	Penn TreeBank	Si
Stanford NLP	API	Si	Eagles	Si

Fuente: http://170.210.201.137/pdfs/asai/ASAI-06.pdf

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text, token.pos_)
```

RUN

```
Veo al hombre con el telescopio.
Veo VERB
al ADP
hombre NOUN
con ADP
el DET
telescopio NOUN
. PUNCT
```

[7] Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python. In XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48 (Salta). Consultado en: http://170.210.201.137/pdfs/asai/ASAI-06.pdf

Segmentación

- Algunas palabras son indivisibles: *que, no, ya, y…*
- Pero en general las palabras tienen partes:
  - Raíz o raíces
  - Afijos
    - Flexivos: singular/plural, femenino/masculino, tiempos verbales.
    - Derivativos: prefijos, interfijos, sufijos

CANT-O
CANT-ABA-MOS
CANT-O-S

ABRE-LATA-S
CORRE-CAMINO-S

GAT-A
GAT-IT-O
GAT-IT-OS

AMOR
EN-AMOR-AR
DES-EN-AMOR-AR

CONSTITU-IR
CONSTITU-CIÓN
CONSTITU-CION-AL
CONSTITU-CION-AL-IZ-AR
CONSTITU-CION-AL-IZ-A-CIÓN

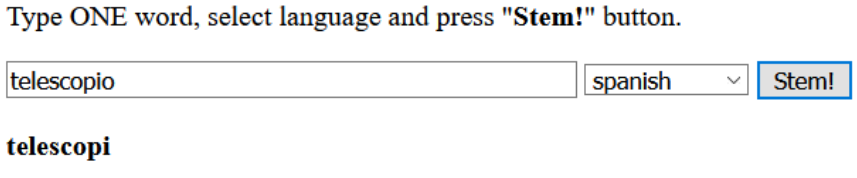
"A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS"

- ¿Cuántas palabras hay en este texto?
- ¿Cuántas palabras diferentes?

Hay 2 tareas diferentes que ayudan a buscar las palabras y sus raíces: stematización y lematización.

Stemmer

- Los algoritmos de stemming intentan reducir las palabras flexionadas y derivadas en su forma raíz, es decir, extraer la raíz de una palabra, la raíz lingüística a la que pertenece; generalmente son algoritmos muy rápidos.
- El stemming es una forma de análisis morfológico.
- Este proceso se realiza porque la raíz de una palabra puede aparecer más veces en un texto.
- El algoritmo más común para stemming es el algoritmo de Porter[8]. Existen además métodos basados en análisis lexicográfico y otros algoritmos similares (KSTEM, stemming con cuerpo, métodos lingüísticos, entre otros).



http://proiot.ru/jssnowball/

[8] https://www.nltk.org/howto/stem.html

```
from nltk import word_tokenize
from nltk.stem.snowball import SnowballStemmer

stemmer = SnowballStemmer("spanish")
doc = "A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS"
' '.join( [stemmer.stem(word) for word in word_tokenize(doc)] )
```

```
'a la gat le gust tra gat que tra a mas gat y a sus gatit'
```

Lematizador

- Lematización de los términos, es una parte del procesamiento lingüístico que trata de determinar el lema de cada palabra que aparece en un texto.
- Su objetivo es reducir una palabra a su raíz, de modo que las palabras clave de una consulta o documento se representen por sus raíces en lugar de por las palabras originales.
- El lema de una palabra comprende su forma básica más sus formas declinadas.
- La lematización requiere etiquetado POS.
- La lematización requiere un diccionario como WordNet o Wikipedia.
- La lematización es un proceso computacionalmente costoso, en comparación con el stemming.

https://spacy.io/models/es

Try out the model

```
import spacy

nlp = spacy.load("es_core_news_sm")
doc = nlp("A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS")
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.lemma_)
```

RUN

```
A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS
A ADP a
LA DET el
GATA NOUN gata
LE PRON él
GUSTA ADP GUSTA
TRAER PROPN TRAER
GATAS NOUN gata
QUE PRON que
TRAEN PROPN TRAEN
A ADP A
MÁS ADV más
GATOS NOUN gatos
Y CCONJ Y
A ADP A
SUS PROPN SUS
GATITOS NOUN gatito
```

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[4])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.lemma_)
```

RUN

```
El gato come pescado.
El DET el
gato PROPN gato
come VERB come
pescado ADJ pescado
. PUNCT .
```

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.lemma_)
```

RUN

```
Veo al hombre con el telescopio.
Veo VERB ver
al ADP al
hombre NOUN hombre
con ADP con
el DET el
telescopio NOUN telescopio
. PUNCT .
```

Analizador de dependencias

- El análisis de dependencia extrae un árbol sintáctico de un texto dado.
- Los árboles representan las relaciones sintácticas entre las palabras de una oración.
- El análisis utiliza un corpus marcado manualmente.
- El análisis requiere más recursos computacionales que los anteriores.

https://spacy.io/models/es

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

RUN

```
Veo al hombre con el telescopio.
Veo VERB ROOT
al ADP case
hombre NOUN obj
con ADP case
el DET det
telescopio NOUN obl
. PUNCT punct
```

Ejercicio3(es)-PLN.ipynb

Tarea:

1. A partir del texto “El Ramo Azul”, de Octavio Paz, publicado en el libro español “Arenas movedizas” en 1949.
   “El ramo azul” trata de un viajero que pasa una noche en un pueblo inquietante. La trama se centra en el diálogo que ocurre cuando un hombre débil se acerca al narrador para intentar sacarle los ojos. Lo que expone Paz, con los ojos como un símbolo, es que hay límites para la percepción.
   A) ¿Cuántas palabras hay en el texto?
   B) ¿Cuántas palabras diferentes existen?
   C) ¿Qué cantidad de sustantivos, adjetivos y verbos posee el texto?
- Para el inciso C) puede tomarse en cuenta la ayuda de otras herramientas implementadas. Ejemplo: el servicio FreeLing: http://www.corpus.unam.mx/servicio-freeling/

Conclusiones

- El PLN es fácil de entender, posible y tiene gran importancia en nuestra época de información.
- La comprensión total del lenguaje natural es un objetivo aún distante. Pero existen herramientas y sistemas útiles para resolver varios problemas prácticos de PLN. El reto consiste en encontrar la correcta adecuación entre los diferentes tipos de problemas y las herramientas disponibles para resolverlos.