UNAM iimas INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Etiquetado y procesamiento de texto

Materia: Procesamiento de lenguaje natural

Blanca Vázquez

Categorías gramaticales

En el nombre bajo el que se agrupan todas las palabras de una lengua, distribuidas por clases.

Estas clases son: sustantivos, pronombres, adjetivos, artículos, verbos, preposiciones,

Imagen tomada de https://liec.dgb.unam.mx/index.php/escribe/la-gramatica-lo-mas-basico/las-clases-de-palabras

Categorías gramaticales

Categoría léxica o de contenido	Categoría funcional
Sustantivo Adjetivo Verbo	Pronombre Determinante Adverbio Preposiciones Conjunción

Categorías gramaticales

Clase	Descripción	Ejemplo
Sustantivo	Seres, objetos, ideas	1. contable (coche) / incontable (leche) 2. propio (Juan) / común (pan) 3. simple (puerta) / compuesto (lavacoches) 4. concreto (almacén) / abstracto (belleza)
Pronombre	Sustituye al sustantivo	1. Personales: yo, tú, él , nosotros, vosotros, ellos: me, te, se, nos, os, lo, mi, ti, si, le, lo, la... 2. Demostrativos: este, ese, aquel, estos, esos, aquellos... 3. Indefinidos: nada, todo, algo, nadie, alguien, alguno, bastantes, varios, cualquier, cualquiera, cualesquiera... 4. Numerales: un, dos, tres, primero, segundo... 5. Relativos: que, quien, cuyo, cual, cuantos... 6. Posesivos: mío, tuyo, suyo, nuestro, vuestro, suyo... 7. Interrogativos: qué, quién, cuánto, cuándo, cuál, dónde, cómo...
Adjetivo	Cualidades del sustantivo	Grados del adjetivo: 1. positivo: Este es un postre dulce. 2. comparativo: este postre es más dulce que aquel. 3. superlativo: este es un postre muy dulce / dulcísimo.
Verbo	Acciones	1. Simples: Presente, Pretérito imperfecto, Pretérito perfecto simple, Futuro imperfecto, Condicional... 2. Compuestos: Pretérito perfecto compuesto, Pretérito anterior, Futuro perfecto, Pretérito pluscuamperfecto, Condicional perfecto...

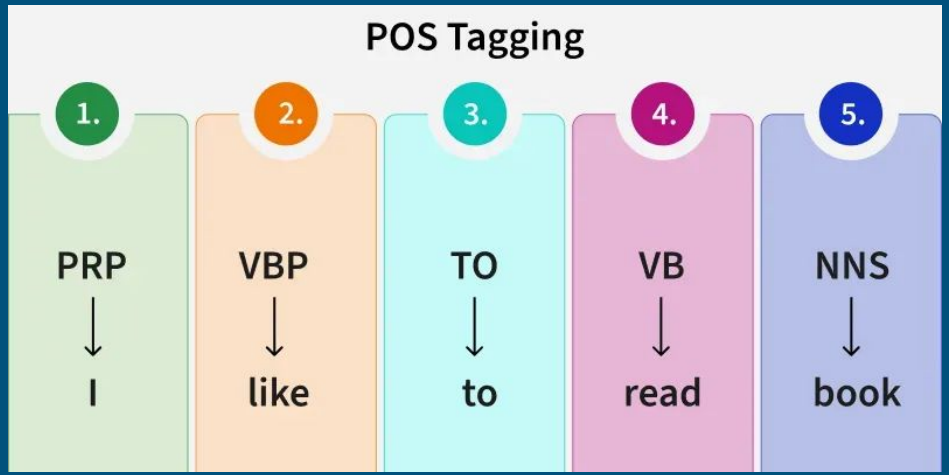
Categorías gramaticales

Clase	Descripción	Ejemplo
Adverbio	Complementan al verbo	1. lugar: lejos, cerca, aquí, allí, allá, acá… 2. modo: así, bien, mal, etc. 3. tiempo: ayer, mañana, nunca, hoy, jamás, siempre, a veces. 4. duda: quizás, tal vez, acaso. 5. cantidad: mucho, poco, bastante, demasiado. 6. afirmación: sí, también. 7. negación: no, tampoco.
Determinante	Acompañan al sustantivo determinandolo	1. Artículos: El, la, los, las, un, una, unos, unas. 2. Demostrativos: Este, ese, aquel, etc. 3. Posesivos: Mi, tu, su, etc. Indican pertenencia. 4. Numerales: Uno, dos, tres, primero, segundo, etc. 5. Indefinidos: Alguno, ninguno, mucho, poco, etc. 6. Interrogativos y Exclamativos: ¿Qué?, ¿Cuál?, ¡Qué!, ¡Cuánto!. 7. Relativos: Cuyo, cuya, cuyos, cuyas.
Preposición	Enlazan palabras	ante, bajo, cabe, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre, tras.
Conjunción	Enlazan palabras y oraciones	1. Copulativas: y, e, ni, 2. Disyuntivas: o, u. 3. Adversativas: pero, mas, sino. 4. Concesiva: aunque. 5. Causales: porque, pues, 6. Condicionales: si. 7. Comparativa: tan, tanto, que, como …

¿Cuál es la importancia / utilidad de dar a cada palabra una categoría en el procesamiento de lenguaje natural?

Desambiguador morfosintáctico

Un desambiguador morfosintáctico, conocido como POS tagger (Parts-Of-Speech), es una herramienta para asignar etiquetas gramaticales a cada palabra dentro de un texto.



Las etiquetas POS ayudan a comprender la función y el significado de las palabras dentro de un contexto

Imagen tomada de *geeksforgeeks*, 2015

Flujo de trabajo de etiquetado POS

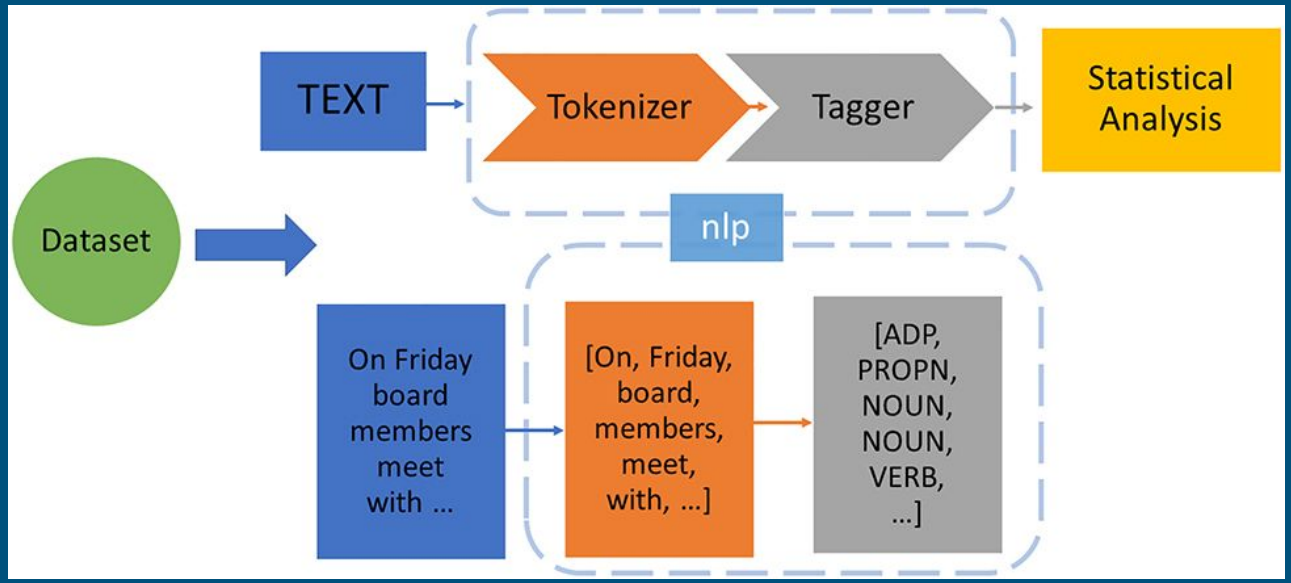


Imagen tomada Mendhakar, 2020.

Etiquetado Penn Treebank

Es una lista de etiquetas de partes del discurso, es decir, las etiquetas indican la parte del discurso y, a menudo también otras categorías gramaticales (caso, tiempo, etc.) de cada token en un corpus de texto .

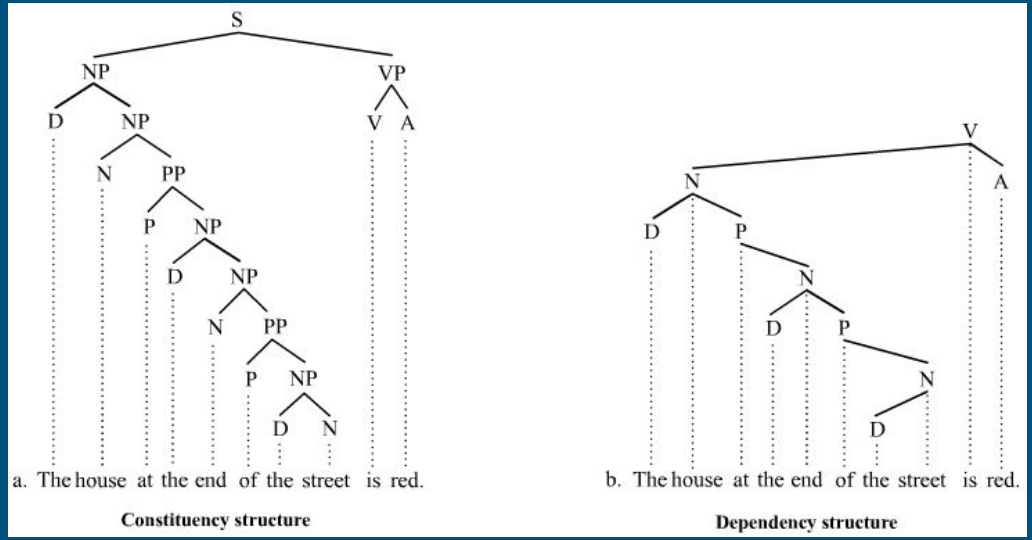


Imagen tomada de https://en.wikipedia.org/wiki/File:The_house_at_the_end_of_the_street.jpg

Etiquetado Penn Treebank

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

Imagen tomada de https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Etiquetado: Universal dependencies

- *Universal dependencies* (UD) proporciona un inventario universal de categorías para facilitar la anotación consistente en diferentes idiomas, permitiendo a la vez extensiones específicas de cada idioma cuando es necesario.
- Facilita el desarrollo de analizadores multilingües.
- El esquema de anotación se basa en una evolución de las dependencias universales de:
  - Stanford (de Marneffe et al., 2006, 2008, 2014),
  - Categorías gramaticales de Google (Petrov et al., 2012)
  - Interlingua *Interset* para conjuntos de etiquetas morfosintácticas (Zeman, 2008).

Etiquetado: Universal dependencies

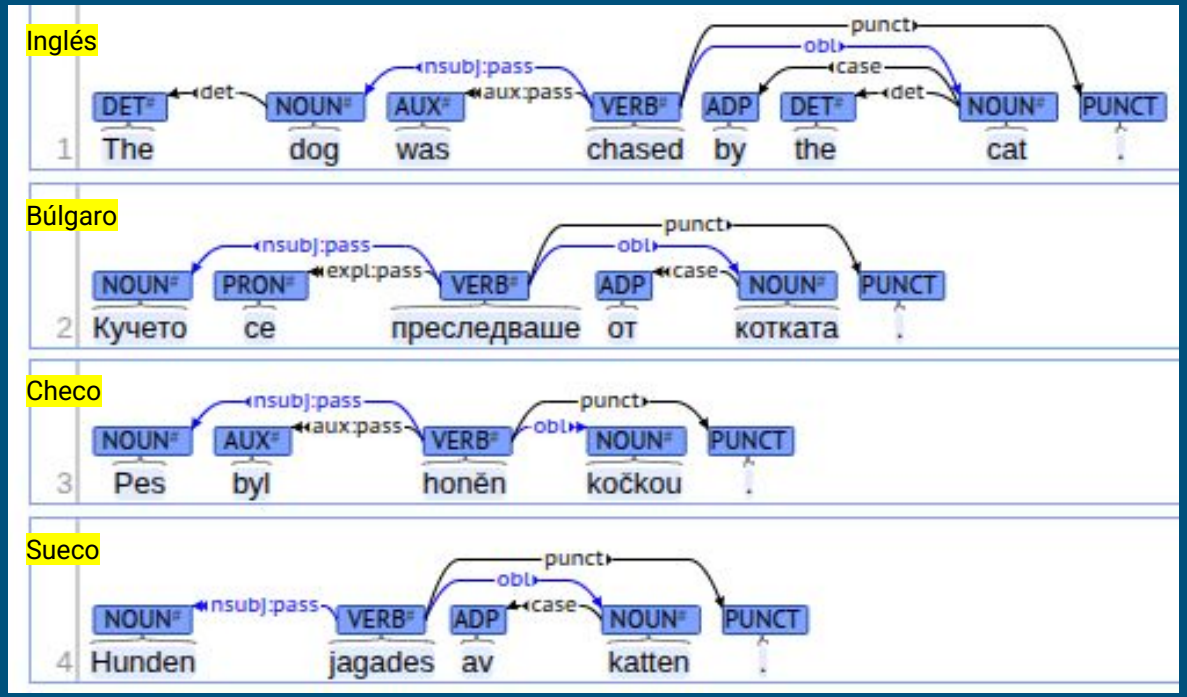


Imagen tomada de https://universaldependencies.org/introduction.html

Herramientas para etiquetado y procesamiento de texto

Natural Language Toolkit

spaCy

TreeTagger

Institute for Computational Linguistics of the University of Stuttgart.

FreeLing

Universitat Politècnica de Catalunya

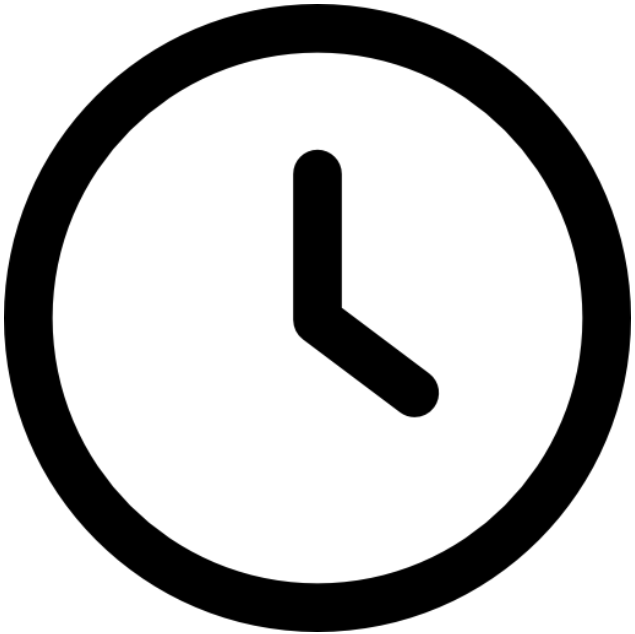
Stanza

Stanford University

CitiusTagger and CitiusNec

Universidad de Santiago de Compostela

Time to code



Reconocimiento de entidades nombradas

El Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés) es una tarea de PLN que identifica y clasifica entidades nombradas, como personas, organizaciones, ubicaciones, fechas, entre otras.

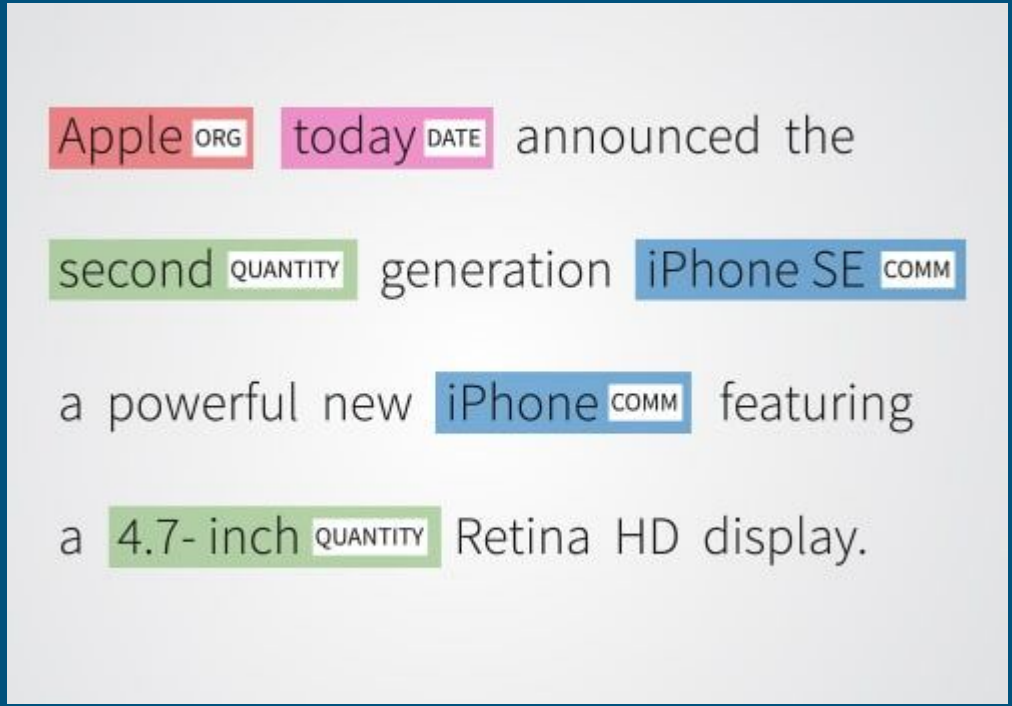


Imagen tomada de https://es.shaip.com/blog/named-entity-recognition-and-its-types/

Entidades nombradas más comunes

Named Entity Type	Example
ORGANIZATION	WHO
PERSON	President Obama
LOCATION	Mount Everest
DATE	2020-07-10
TIME	12:50 P.M.
MONEY	One Million Dollars
PERCENT	98.24%
FACILITY	Washington Monument
GPE	North West America

Imagen tomada de https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/

Ventajas

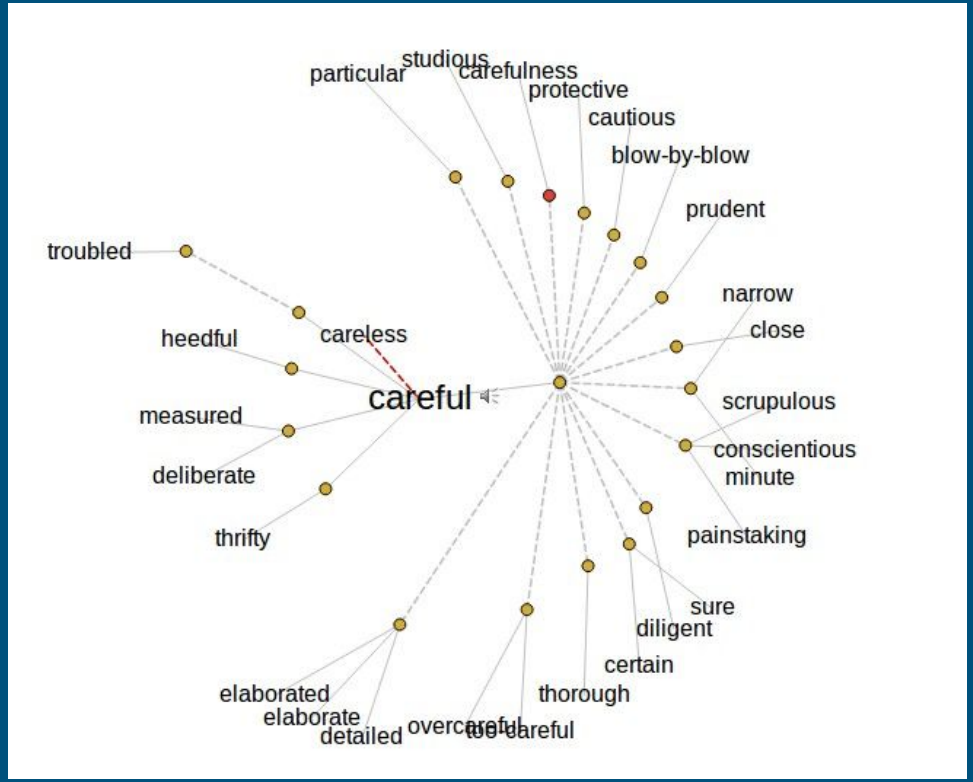
Ventajas	Descripción
Simplificación de texto	Ayuda a deconstruir oraciones complejas para una comprensión más fácil.
Recuperación de información mejorada	Permite una indexación y búsqueda más precisa basadas en categorías gramaticales.
Named Entity Recognition (NER)	Sirve como precursor para identificar nombres, lugares y organizaciones.
Análisis sintáctico	Ayuda a analizar la estructura de las oraciones y las relaciones entre palabras.

Desventajas

Ventajas	Descripción
Ambigüedad	Las palabras pueden tener múltiples significados dependiendo del contexto.
Expresiones idiomáticas	Las frases informales o no estándar son difíciles de etiquetar correctamente.
Palabras fuera de vocabulario	Las palabras no vistas pueden provocar un etiquetado incorrecto.
Dependencia del dominio	Es posible que los modelos no se generalicen bien fuera de su dominio de entrenamiento

WordNet

- WordNet es una base de datos léxica del Idioma inglés que agrupa palabras conjuntos de sinónimos llamados *synsets*.
- Proporciona definiciones cortas y generales.
- Almacena relaciones semánticas entre los conjuntos de sinónimos.
- Produce una combinación de diccionario y tesauro.
- WordNet es un lexicón computacional (desambiguar el significado de las palabras: *word sense disambiguation* WSD.
- Asigna el concepto más apropiado (i.e. synsets) a los términos en contexto.



Para mayor información: https://wordnet.princeton.edu/

WordNet como una ontología

```
dog, domestic dog, Canis familiaris
   => canine, canid
      => carnivore
        => placental, placental mammal, eutherian, eutherian mammal
          => mammal
            => vertebrate, craniate
              => chordate
                => animal, animate being, beast, brute, creature, fauna
                  => ...
```

Imagen tomada de https://es.wikipedia.org/wiki/WordNet