3. ANÁLISIS LÉXICO

Contenido

3.1 ¿Qué es el análisis léxico?

- El léxico es el conjunto de palabras de una lengua.
- En PLN el análisis léxico se centra en el reconocimiento y categorización de las palabras.
- La definición de palabra que se ocupará para este curso será: *Todo aquello que se encuentre entre dos espacios en blanco*.
- Las palabras en este caso serán identificadas y etiquetadas según sus categorías gramaticales.

3.2 Categorías gramaticales

Las categorías gramaticales también se denominan clases de palabras

Las palabras pueden clasificarse de diferentes formas según diversos criterios. Una de las formas en las que se hace esta clasificación es mediante su función gramatical.

Por ejemplo, en la obra literaria de Julio Cortázar Rayuela

"Apenas él le **amalaba** el **noema**"

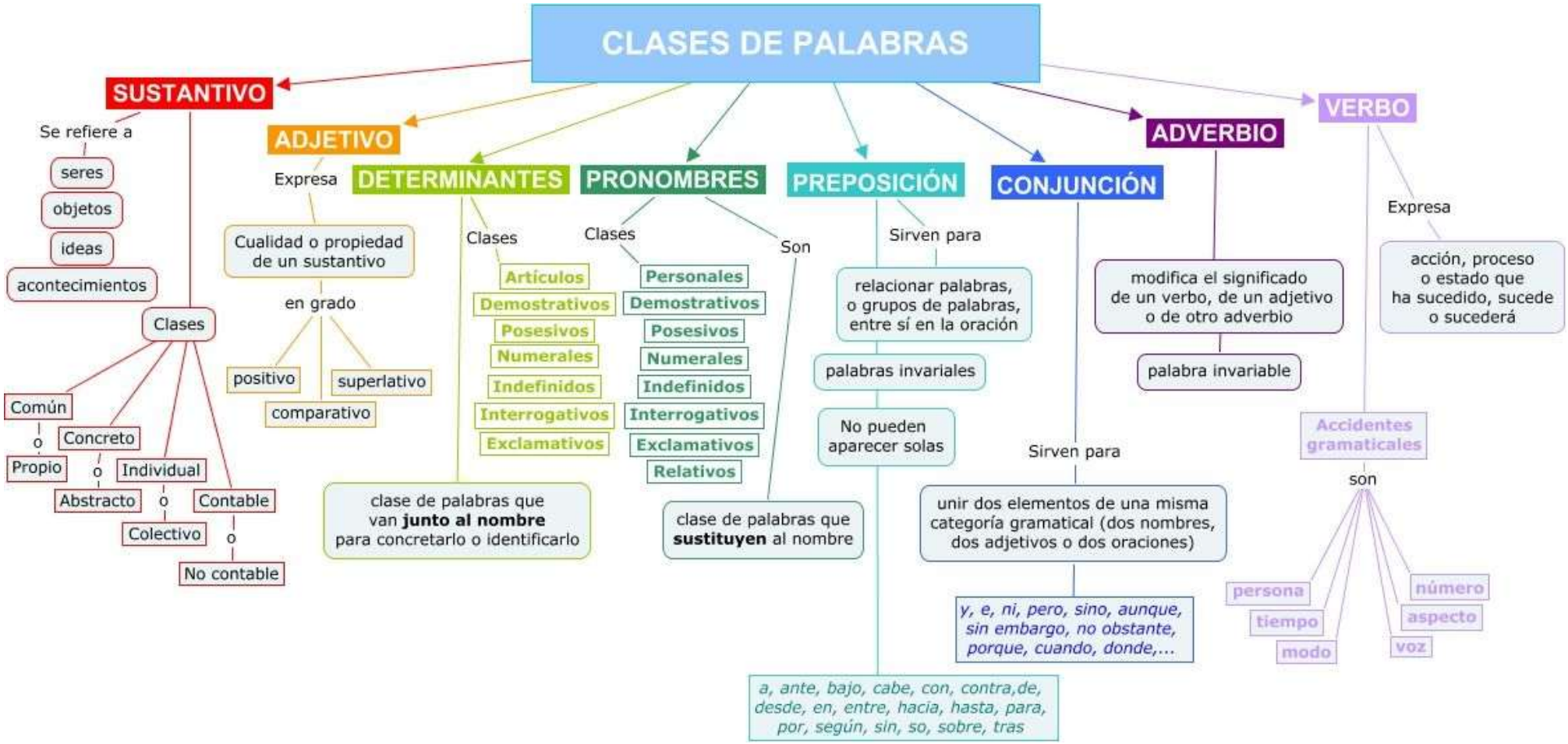
No sabemos que significan las palabras en negritas, pero es posible intuir que 'amalaba' aparece en construcciones como 'el amaló', 'ellas amalaban' o 'no me gustó su forma de amalar'.

Con base en lo anterior podemos vislumbrar que 'amalaba' es una clase particular de palabra, por lo tanto podemos inferir sus posibilidades de contexto de aparición y complementos. Es a partir de sus restricciones de aparición y función que se dividen las categorías gramaticales.

3.2 Categorías gramaticales

Las categorías gramaticales son las siguientes:

- Sustantivo o nombres (N) *Harry, niño, clima, policía*
- Verbo (V) *arribar, discutir, derretir, oír, recordar*
- Adjetivo (A) *bueno, alto, silencioso, viejo, costoso*
- Preposición (P) *en, de, a, con, desde, entre, hacia*
- Adverbio (Adv) *silenciosamente, lentamente, ahora*
- Determinantes (Det) *la, eso, aquello, ese, un, algún, ningún*
- Conjunción (Conj) *y, o, pero, cuando, porque, es decir*
- Pronombres (Pron) *lo, mi, me, ella, él, le, te*


CLASES DE PALABRAS
SUSTANTIVO
Se refiere a
seres
objetos
ideas
acontecimientos
Clases
Común
o
Propio
Concreto
o
Abstracto
Individual
o
Colectivo
Contable
o
No contable
ADJETIVO
Expresa
Cualidad o propiedad de un sustantivo
en grado
positivo
comparativo
superlativo
DETERMINANTES
Clases
Artículos
Demostrativos
Posesivos
Numerales
Indefinidos
Interrogativos
Exclamativos
clase de palabras que van **junto al nombre** para concretarlo o identificarlo
PRONOMBRES
Clases
Personales
Demostrativos
Posesivos
Numerales
Indefinidos
Interrogativos
Exclamativos
Relativos
Son
clase de palabras que **sustituyen** al nombre
PREPOSICIÓN
Sirven para
relacionar palabras, o grupos de palabras, entre sí en la oración
palabras invariales
No pueden aparecer solas
a, ante, bajo, cabe, con, contra,de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre, tras
CONJUNCIÓN
Sirven para
unir dos elementos de una misma categoría gramatical (dos nombres, dos adjetivos o dos oraciones)
y, e, ni, pero, sino, aunque, sin embargo, no obstante, porque, cuando, donde,...
ADVERBIO
modifica el significado de un verbo, de un adjetivo o de otro adverbio
palabra invariable
VERBO
Expresa
acción, proceso o estado que ha sucedido, sucede o sucederá
Accidentes gramaticales
son
persona
tiempo
modo
número
aspecto
voz

3.3 Conteo de palabras: Tokenización

En PLN se puede realizar el reconocimiento de palabras por clase específica a este proceso se le conoce como **tokenización**. En esta segmentación las palabras reconocidas reciben el nombre de Token.

- Entonces, un token o palabra es cada una de las formas que aparecen en el texto, sin importar cuántas veces ocurra cada una.
- En cambio, un type o tipo es cada una de las formas o palabras diferentes que aparecen en un texto

  EL CAMIÓN TOMÓ EL DESVÍO PARA TOMAR LA VÍA RÁPIDA

  tokens= 10, types=?

Los ítems obtenidos de este proceso sirven para su análisis estadístico, lingüístico y para recolección de datos que pueden ser recuperados para posteriores análisis.

Conteo de palabras: Tokenización

Ejemplo:

Podría ponerme a escribir la tesis, **podría**…

7 token, 6 types

*En este caso el analizador no segmenta diferenciando los tokens por mayúsculas y minúsculas, sin embargo existe la posibilidad de hacerlo.

TOKEN ≥ TYPE

3.3 Conteo de palabras: listas de palabras

En el análisis léxico, después de segmentar el texto, de tokenizarlo, es necesario llevar los resultados de este a un nivel cuantitativo que permita hacer estudios posteriores de comportamiento de la lengua.

En este caso se hacen **listas de palabras**, las cuales pueden ser simples, de formas canónicas, de lemas, de dos o más palabras, o de partes de la oración.

3.3 Conteo de palabras: listas de palabras

Tipos de listas:

- **Lista simple**: muy utilizada, se presenta una lista de los types y su frecuencia.
- **Lista de formas canónicas**: se presentan las entradas de diccionario de forma alfabética y a cada línea le suceden las palabras que le corresponden a dicha forma canónica.
- **Lista de lemas**: parecida a la anterior, se presenta el lema o la raíz y después las palabras que le corresponden. En esta lista se admiten palabras de diferente categoría gramatical que el lema.
- **Lista de dos o más palabras**: se presentan grupos de dos o más palabras que ocurren contiguamente (sería este el caso de las colocaciones, que veremos más adelante). Su importancia radica en la localización de unidades multiléxicas.
- **Lista de partes de la oración**: es una lista de las palabras en la que se indica la categoría gramatical a la que pertenece, esto con fines estadísticos.

3.3 Conteo de palabras: frecuencias

Como mencionamos, en estas listas se suele poner la forma o palabra y su frecuencia. Existen dos tipos de frecuencia:

- **Frecuencia absoluta**: es el número de veces que se repite una palabra en el corpus.
- **Frecuencia relativa**: es el número de veces que aparece una palabra dentro del texto con relación al total de palabras del corpus. Su valor es igual a la frecuencia absoluta entre la suma de frecuencias absolutas (total de tokens).

  $Fr = Fa / \Sigma Fa$

3.4 Colocaciones

- Las **colocaciones** son *combinaciones de dos o más palabras que ocurren en la lengua con más frecuencia de lo normal*.
- En PLN suele referirse a la ocurrencia en la que las palabras pueden estar acomodadas en el texto y las condiciones en las que ocurre la cercanía de algunas palabras en contextos determinados.
- El estudio de las colocaciones permite observar patrones de coocurrencia de palabras, es decir, bajo qué circunstancia suele aparecer una palabra cerca de otra.
- La distancia de las colocaciones puede ir desde cero, una palabra contigua a la otra, hasta una distancia relativamente corta, por ejemplo, 6 palabras de distancia una de otra.

3.4 Colocaciones

Tipos de colocaciones:

- Nombres de organismos: Instituto Nacional Electoral, Universidad Nacional Autónoma de México, Organización de las Naciones Unidas.
- Expresiones terminológicas: labio leporino, inteligencia artificial, procesamiento de lenguaje natural.
- Expresiones discursivas: por favor, cómo te va, sin embargo, vamos a ver.
- Locuciones (nominales, adjetivales, verbales, etc.): blanco y negro, ojo de buey, de las mil maravillas, hacer caso, llevar a cabo, a pesar de.
- Expresiones idiomáticas: dar gato por liebre, empinar la jarra, doblar las manitas.

3.4 Colocaciones

Elementos de una colocación:

- Nodo (base, palabra clave/llave): la palabra de la cual se busca la colocación.
- Colocativo (correlativo léxico): cualquier palabra que coocurre con el nodo.
- Ventana: contexto en que coocurren los colocativos.

3.4 Colocaciones

La cuantificación de las colocaciones toma en cuenta los siguientes elementos:

- Conteo de palabras cercanas
- Conteo de las palabras que aparecen hacia la derecha
- Conteo de las palabras que aparecen hacia la izquierda
- Conteo de las concordancias

3.4 Colocaciones

Información mutua

- Para obtener la fuerza de asociación, la reducción de entropía y la información de una variable aleatoria sobre otra, se utiliza el algoritmo de información mutua (MI):

$$MI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

- La información mutua entre dos palabras *x* e *y* se expresa como *MI(x, y)* o como *I(x, y)*, y corresponde al logaritmo base dos del cociente de la probabilidad de que las dos palabras ocurran juntas entre la probabilidad de ocurrencia de cada una de las palabras en el corpus. A mayor MI mayor fuerza de asociación.

3.4 Colocaciones

Criterio absoluto

- Se usa para determinar cuál es la colocación en expresiones que pueden tener más de dos palabras. Ejemplo: 1) *un número*, 2) *un número de*, 3) *un número de veces*. Se duda si la colocación es 1) o 2).
- Para esto se usa la fórmula del criterio absoluto, que determina los costos de una colocación (K). La expresión de mayor costo es la colocación.

$$K(a) = (|a| - 1) \times (f(a) - f(b))$$

a = colocación

|a|= longitud de la colocación a

f(a)= la frecuencia de ocurrencia de la colocación a

f(b)= la frecuencia de la colocación siguiente en cuanto al número de palabras

Colocaciones

- Por ejemplo, si ”*un número*" ocurre 102 veces en un corpus, ”*un número de*" ocurre 51 veces, y ”*un número de veces*" ocurre 20 veces, entonces:

K(*un número*) =

(/*un número*/ -1) (f(*un número*) - f(*un número de*) =

(2 -1) (102 - 51) = 1 x 51 = 51

K(*un número de*) =

(/*un número de*/ -1) (f(*un número de*) - f(*un número de veces*) =

(3 -1) (51 - 20) = 2 x 31 = 61

Por tanto, *un número de* es la colocación.

3.5 N-gramas

- Uno de los principales objetivos que persigue el procesamiento del lenguaje natural es que sea posible representar los datos recolectados en información cuantitativa.

- Un n-grama es la secuencia ordenada de n elementos. Si n=2 se denominan bigramas; n=3, trigramas; para n>=4 entonces se llaman Modelos de Markov de orden (n-1).

- La representación de las secuencias funciona para hacer las determinaciones estadísticas, pero también para la elaboración de predictores de texto.

3.5 N-gramas

Dada una secuencia S= $s_1 s_2 s_3 ... s_k$, se denomina n-grama a cualquier subsecuencia, A=$s_{i+1} s_{i+2} ... s_{i+n}$, donde *i* es un valor entre 0 y $|S|-n$ para garantizar que la longitud de *A* sea siempre *n* o lo que es lo mismo

$$|A|=n;\ n>1$$

Estas secuencias se pueden convertir en representaciones estadísticas, ya sea visuales o de patrones.

3.6 Lexicón computacional

- Un **lexicón** es una colección de palabras, de las cuales se conoce su variabilidad léxica (prefijos y sufijos, raíces y otras formas o variaciones). Suele ser la base de trabajo para el PLN.
- Los lexicones aportan principalmente información sintáctica y semántica, lo que permite tratar la sinonimia, la antonimia y la polisemia, esta información se hace evidente mediante el etiquetado.
- Entonces, un lexicón es básicamente una gran base de datos con palabras que guardan relaciones entre sí (ya sea sintácticas o semánticas). Un ejemplo, es WordNet.

3.6 Lexicón computacional: WordNet

- **WordNet®** es una gran base de datos léxica (en Inglés). Las palabras se agrupan en conjuntos de sinónimos cognitivos (synsets). Los ***synsets*** se interconectan mediante relaciones semánticas y léxicas.
- WordNet se diferencia de un thesaurus en:
  1. WordNet relaciona no solo formas léxicas, sino sentidos específicos de las palabras. Así, se ayuda a la desambiguación de los significados.
  2. WordNet etiqueta relaciones semánticas entre palabras, mientras que los thesaurus no especifican los tipos de relaciones.

- https://wordnet.princeton.edu/

3.6 Lexicón computacional: WordNet

- WordNet se estudiará con más detenimiento cuando hablemos del análisis semántico.
- Herramientas como esta son de gran ayuda para realizar tanto análisis léxico, como sintáctico y semántico.
- Por ahora pueden visitar el sitio y hacer algunas consultas para que vayan familiarizándose con la forma en que se puede organizar la información lingüística dentro de un lexicón.

Bibliografía

Mijangos, V. (2017) Procesamiento de Lenguaje Natural. https://sites.google.com/site/victormijangoscruz/cursos/procesamiento-de-lenguaje-natural

Moreno-Ortiz, A., 1998. El lexicón en la lexicografía computacional: adquisición y representación de la información léxica.

Jurafsky, Daniel & James H. Martin. (2008). Speech and Language Processing: *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (2nd edition), Prentice Hall.