

Tema 1. Procesamiento de Lenguaje Natural

Tecnologías Emergentes en la Sociedad de la Información

Motivación



- Exponencial crecimiento y disponibilidad de datos.
 - 2013, 3.5 ZB
 - 2022, 40 ZB
 - 2025, 180 ZB
 - Más de un 80%, datos no estructurados y en formato texto

1 ZB = 1 trillón de GB

Procesamiento de Lenguaje Natural (PLN)

- Objetivo: crear aplicaciones informáticas capaces de entender y/o generar lenguaje humano.

Campo multidisciplinar

- Lingüística: modelos formales del lenguaje.
- Informática y Matemáticas: estructuras de datos y algoritmos, modelos probabilísticos, programación dinámica, clasificadores (aprendizaje automático), autómatas, etc.
- Psicología

Aplicaciones PLN

- Sistemas de Recuperación de Información

The screenshot shows a Google search results page. The search query is "estudios máster ciencia y tecnología informática". The results are filtered under the "Todo" tab. The first result is a link to the UC3M website for their Master's program in Science and Technology Informatics. The second result is a link to the University of Valencia's website for their Double Master's program in Computer Engineering and Science and Technology. The third result is a link to the University of Valencia's website for their Master's program in Science and Technology Informatics. The fourth result is a link to the UPM website for their Master's program in Sciences and Technologies of Computing. The fifth result is a link to the UCM website for their Master's program in Informatics.

Google estudos máster ciencia y tecnología informática

Todo Noticias Vídeos Imágenes Maps Más Configuración Herramientas

Aproximadamente 9.360.000 resultados (0,75 segundos)

Máster Universitario en Ciencia y Tecnología Informática | UC3M
https://www.uc3m.es/master/ciencia-tecnologia-informatica ▾
El Máster Universitario en Ciencia y Tecnología Informática ofrece una formación avanzada diferencial y puntera que capacita a nuestros alumnos con una ...

Doble Máster en Ingeniería Informática y Ciencia y Tecnología ...
https://www.uc3m.es/master/doble-ingeneria-informatica-ciencia-tecnologia-informati... ▾
MÁSTER EN CIENCIA Y TECNOLOGÍA INFORMÁTICA ... en una posición inmejorable para tu inserción laboral y/o el acceso a los estudios de Doctorado.

Máster Universitario en Ciencia y Tecnología Informática. Universidad ...
www.universia.es › Estudios universitarios ▾
Máster Universitario en Ciencia y Tecnología Informática. Conoce toda la información y datos de interés que necesitas saber sobre este estudio en Universidad ...

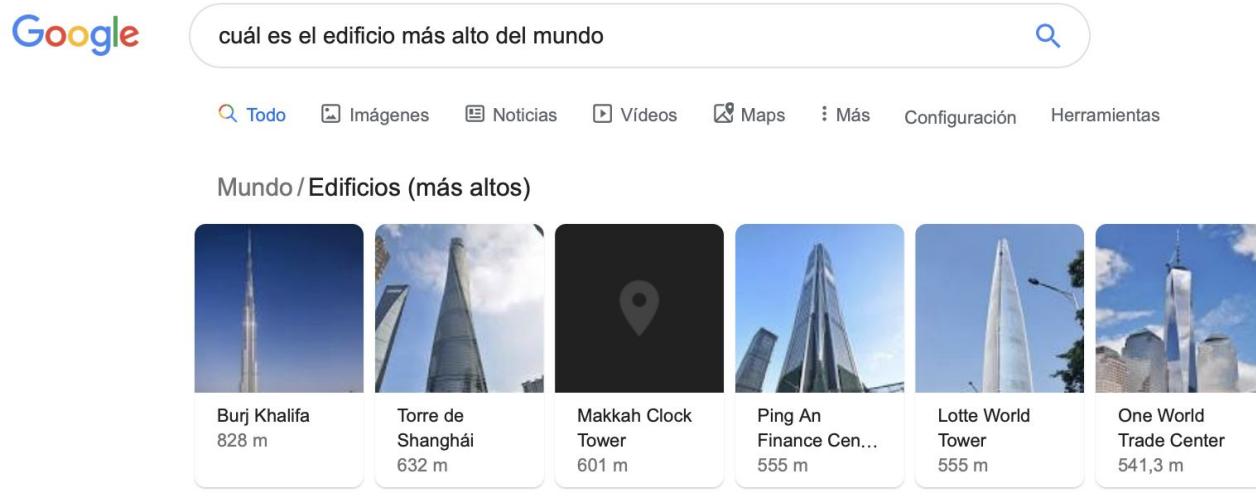
Máster Universitario en Ciencias y Tecnologías de la Computación ...
https://www.etsisi.upm.es/estudios/master/mctcig ▾
El máster en Ciencias y Tecnologías de la Información es un máster que tiene ... Estas ramas de la Informática, disciplinas de I+D+i sólidamente asentadas en ...

Títulos oficiales de Máster -ETS de Ingenieros Informáticos (UPM)
https://www.fi.upm.es/es/masteroficiales ▾
11 abr. 2019 - Consulta de precios públicos para estudios de postgrado, curso 2018-2019 ... Máster Universitario en Ingeniería Informática (Profesión de Ingeniero en ... de la Ciencia y la Tecnología, un mayor grado de conocimientos en ...

Máster - Facultad de Informática
https://informatica.ucm.es/master ▾
Máster. Másters ofertados, acceso y matrícula a los estudios. Instrucciones generales de matrícula de Máster ... Ciencias Sociales y Jurídicas. ... Coordinador de la especialidad de Informática y Tecnología del Máster en Formación del ...

Aplicaciones PLN

- Sistemas de Recuperación de Información



[¿Cuál es el edificio más alto del mundo? | Plataforma ...](#)

<https://www.plataformaarquitectura.cl> › ArchDaily › Artículos ▾

23 ene. 2019 - En la actualidad existen instituciones especializadas que establecen los parámetros para definir objetivamente cuánto mide un **edificio**.

Aplicaciones PLN

- Sistemas de Recuperación de Información

The screenshot shows a search results page from PubMed. At the top, there is a search bar with the query "drug drug interactions". Below the search bar are buttons for "Create RSS", "Create alert", and "Advanced". The search parameters are set to "Format: Summary", "Sort by: Most Recent", and "Per page: 20". On the right side, there is a "Send to" button. The main content area is titled "Best matches for drug drug interactions:" and lists several research articles:

- [Intestinal Drug Interactions Mediated by OATPs: A Systematic Review of Preclinical and Clinical Findings.](#)
Yu J et al. J Pharm Sci. (2017)
- [Pharmacokinetic drug-drug interactions in the intensive care unit - single-centre experience and literature review.](#)
Łój P et al. Anaesthesiol Intensive Ther. (2017)
- [\[Drug-Drug Interactions with Consideration of Pharmacogenetics\].](#)
Ozawa S et al. Yakugaku Zasshi. (2018)

A blue button at the bottom of this list says "Switch to our new best match sort order".

Below this, the section "Search results" is labeled. It shows "Items: 1 to 20 of 287111" and a page navigation bar with buttons for << First, < Prev, Page 1 of 14356, Next >, and Last >>. The first item listed is:

[CROI 2019: advances in antiretroviral therapy.](#)
1. Taylor BS, Tieu HV, Jones J, Wilkin TJ.
Top Antivir Med. 2019 Apr;27(1):50-68.
PMID: 31137003
[Similar articles](#)

The second item listed is:

[Development of calcific aortic valve disease: Do we know enough for new clinical trials?](#)
2. Kostyunin AE, Yuzhalin AE, Ovcharenko EA, Kutikhin AG.
J Mol Cell Cardiol. 2019 May 25. pii: S0022-2828(19)30101-4. doi: 10.1016/j.yjmcc.2019.05.016. [Epub ahead of print]
Review.
PMID: 31136747
[Similar articles](#)

Aplicaciones PLN

- Traducción automática (Machine translation MT).

The screenshot shows the Google Translate interface. At the top, there's a navigation bar with the Google logo, a grid icon, and a "Sign in" button. Below it, the word "Translate" is written in red, along with a "Turn off instant translation" link and a star icon.

The main interface consists of two horizontal language selection bars. The left bar has "English", "Spanish", "French", and "German - detected" with a dropdown arrow. The right bar has "English", "Spanish", "Arabic", and a dropdown arrow, followed by a blue "Translate" button.

On the left, there is a text input box containing German text:

Wir danken all unseren Gästen für die freundlichen Worte und kritischen Anregungen. Jede Anmerkung nehmen wir zum Anlass, unseren eigenen Qualitätsanspruch ständig neu zu prüfen und uns zu verbessern. Wir freuen uns, Sie in unserem Haus begrüßen zu dürfen!

On the right, the English translation is displayed:

We thank all our guests for the kind words and critical suggestions. We take every note as an occasion to constantly revise our own quality requirements and improve ourselves. We are looking forward to welcome you in our house!

Below the English text are several small icons: a star, a square, a speaker, and a left arrow. In the bottom right corner of the translation box, there is a small edit pen icon.

Análisis de Sentimiento



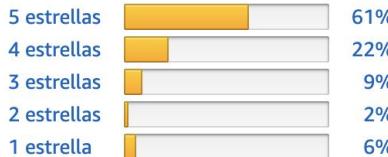
Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón - ... > [Opiniones de clientes](#)

Opiniones de clientes

★★★★★ 4,3 de 5 ▾



620 valoraciones de clientes

[Escribir una opinión](#)

Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón - Pares - - Talla UE

por [Rainbow Socks](#)Precio: [16,99 € - 24,99 €](#) + Envíos gratis con Amazon Prime

★★★★★ Calcetines para verano

11 de junio de 2018

Tamaño: 39/41 | Color: 12 X Multicolor | [Compra verificada](#)

Al verlos lo primero que pensé son demasiado finos y me van a irritar los pies, pero la verdad he quedado muy contenta con ellos, no me irritan y se lavan muy bien. yo los uso para diario (camino bastante) para hacer deporte no se que tal serán.

Actualización después de 6 meses

Al principio estaba muy contenta, pero después de varios lavados se han deteriorado muy rápidamente.

A 2 personas les ha parecido esto útil

[Útil](#)[Comentar](#)[Informar de un abuso](#)

Anna Nardi Fontanals

★★★★★ Perfectos para caminar

27 de octubre de 2018

Tamaño: 39/41 | Color: 12 X Multicolor | [Compra verificada](#)

Me gustan los colores, la cantidad de calcetines y el precio, pero sobretodo son comodos, se ajustan a mi pie genial y se mantienen fijos en los pies. Camino mucho durante el día y con éstos voy muy cómoda, absorbe bien el sudor y NO me hace rozaduras ni nada. Lo único que son un poco finos y no sé que tal me irán ahora en invierno.

A una persona le ha parecido esto útil

[Útil](#)[Comentar](#)[Informar de un abuso](#)

Extracción de Información

Text in

Brazil ranks number 5 in the list of countries by population.

The term "Ibu Negara" (Lady/Mother of the State) is used for wife of the President of Indonesia.

Game of Thrones is an adaptation of A Song of Ice and Fire, George R. R. Martin's series of fantasy novels. It ranks fourth among the IMDB Top Rated TV Shows

Data out

THE COUNTRIES WITH THE LARGEST POPULATION

China	1	1,388,232,693
India	2	1,342,512,706
United States	3	326,474,013
Indonesia	4	263,510,146
Brasil	5	174,315,386

THE COUNTRY'S FIRST LADIES

Brigitte Macron

- Spouse: Emmanuel Macron, President of France (2017 -)

Melania Trump

- Spouse: Donald J. Trump, U.S. President (2017 -)

Iriana Widodo

- Spouse: Joko Widodo, President of Indonesia (2014 -)

- Also known as: "Ibu Negara" (Lady/Mother of the State)

IMDB TOP RATED TV SHOWS

- 1 Planet Earth II (2016) 9.6.
- 2 Band of Brothers (2001) 9.5.
- 3 Planet Earth (2006) 9.5.
- 4 Game of Thrones (2011) 9.4.
- 5 Breaking Bad (2008) 9.4.

Generación de resúmenes

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.



Simplificación de textos

-
- Normal: *Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.*
- Simple: *Alfonso Perez is a former Spanish football player.*

Aplicaciones PLN

- Muchas más...
 - generación de hipótesis,
 - ayuda a la toma de decisiones,
 - soporte para ensayos clínicos y epidemiológicos.
 -

Niveles del lenguaje

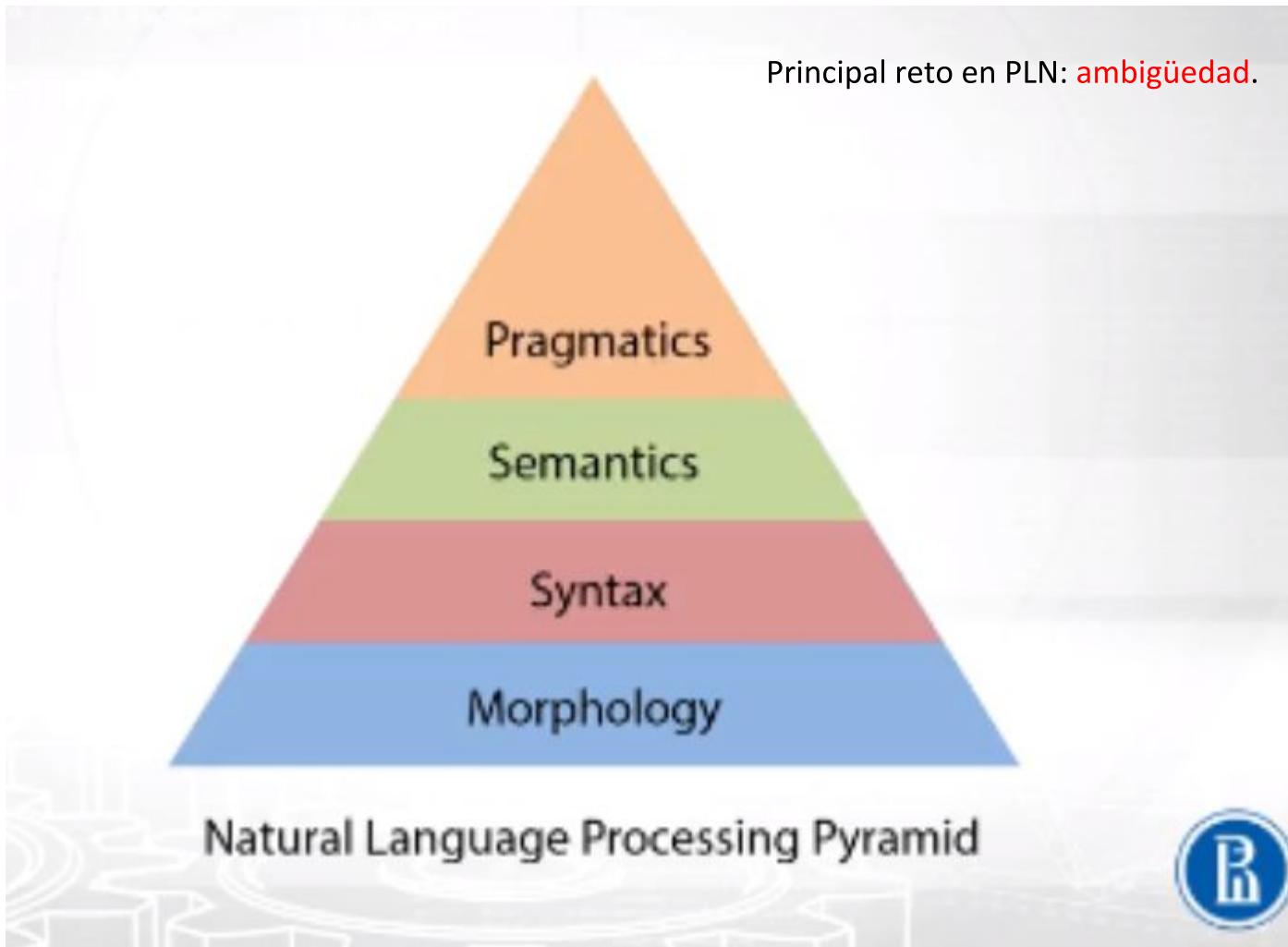


Imagen tomada de <https://www.coursera.org/learn/language-processing>

Ambigüedad en el nivel fonológico

- Inglés:
 - *heel – talón, tacón / heal – sanar, curar*
 - *hear – oír / here – aquí*
 - *I – yo / eye – ojo*
 - *mail – correo / male – masculino*
 - *meet – reunirse, conocer / meat – carne*
- Español:
 - *esconde / es conde*
 - *¿Qué es de Pilar? / ¿Qué es depilar?*
 - *¿Mediste la caja? / ¿Me diste la caja?*

Nivel Morfológico

1. Estudia la composición de las palabras (lexemas y morfemas).
 - Lematización y stemming.
 - Análisis morfológico.

Nivel Morfológico: Stemming

- Obtiene la raíz (stem) de la palabra.
 - *cantariamos* -> *cant-*
 - *cantante* -> *cantant-*
- *Útil en los sistema de búsqueda: poliposis nasal = pólipos nasales*
- Stemmer online: <https://snowballstem.org/demo.html>

Nivel Morfológico: Lematización

- Obtener el lema o forma canónica de una palabra.
 - *cantaremos* -> *cantar*
 - *pequeñito* -> *pequeño*

Lematizador online: <http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm>

Nivel Morfológico: PoS tagging

- **Análisis morfológico:** asignación de la categoría léxica de una palabra en función del contexto en el que aparece.
- Categorías léxicas: Nombre, Verbo, Adjetivo, Adverbio, Pronombre, Determinante, Artículo, Preposición, Interjección.
- Útil en muchas tareas de PLN: reconocimiento de entidades, extracción de relaciones, etc.
- PoS tagger online: <http://nlp.stanford.edu:8080/parser/index.jsp>

Niveles Lenguaje - Ejemplo Análisis Morfosintáctico

Candela juega con sus juguetes nuevos

Palabra	<u>Etiqueta</u> (PoS tag)	Descripción
Candela	NNP	Nombre propio
juega	VBZ	Verbo
con	IN	Preposición
sus	PRP\$	Pronombre Posesivo
juguetes	NNS	Nombre común plural
nuevos	JJ	adjetivo

Penn Treebank Tagset

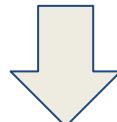
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

CC	Coordinating conjunction	PP\$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential <i>there</i>	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	<i>to</i>
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP\$	Possessive wh-pronoun
PP	Personal pronoun	WRB	Wh-adverb
			Plus additional tags for punctuation.

¿Por qué es útil el análisis morfosintáctico en PLN?

¿Quién inventó la fregona?

PRONOMBRE INTERROGATIVO DE PERSONA

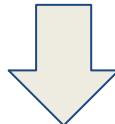


Pista: buscar un nombre propio (NNP)

¿Por qué es útil el análisis morfosintáctico en PLN?

¿Quién inventó la fregona?

PRONOMBRE INTERROGATIVO DE PERSONA



Manuel Jalón Logroño, 31 de enero de 1925 - Zaragoza, 16 de

NNP

NNP

NNP
diciembre de 2011) fue un inventor español. Ingeniero aeronáutico de
formación, y oficial del Ejército del Aire, inventó la fregona y la
NNP

jeringuilla desecharable.

Ambigüedad en el nivel morfológico

- *El chico más **bajo** **bajó** a tocar el **bajo** **bajo** la escalera, **bajo** la dirección del director*
 - (1) bajo - Adjetivo
 - (2) bajó - Verbo (bajar)
 - (3) bajo - Substantivo (instrumento musical)
 - (4) bajo - Preposición (lugar)
 - (5) bajo - Preposición (modo)

Nivel Sintáctico

Estudia las reglas para combinar palabras en sintagmas y oraciones.

- **Tokenización**
- División en oraciones
- Análisis sintáctico (superficial, profundo, de dependencias)

Nivel Sintáctico - Tokenización

- División del texto en tokens.
- Previo al análisis sintáctico y semántico

Nivel Sintáctico - Tokenización

- Una posible estrategia es dividir por espacios en blanco.
- ¿Es correcta siempre?

Aina es buena -> ['Aina', 'es', 'buena']

*Aina es buena, lista y guapa. -> ['Aina', 'es',
'buena,', 'lista,' 'y', 'guapa.] !!!*

La tokenización correcta debería ser:

['Aina', 'es', 'buena', ',', 'lista', ',', 'y', 'guapa', ':']

Nivel Sintáctico - Tokenización

- Y si dividamos por espacios en blanco y signos de puntuación
- ¿Es correcta en todos los casos?
El Dr. aumentó la dosis 13,5 gr. casi un 5%.
IGM S.A. ha ingresados 1.500 millones de euros.
- Debe incluir el tratamiento de signos de puntuación, expresiones numéricas, símbolos, codificación ASCII vs unicode (ej: TFN-alpha,TNF-α).

Nivel Sintáctico - Tokenización

- Contracciones en inglés: *I'm a teacher* -> ['I','m'], ['I' , "am"] , ['I' , "", 'am'] ,
- Hiphenation: *well-known, Ten-year-old, Mother-in-law, co-administration, high-interest*
- Algunos idiomas no tienen espacios entre los tokens (Chino y Japones).
- En algunos idioma (árabe y hebreo), el texto se escribe de derecha a izquierda, excepto los números (de derecha a izquierda)

Nivel Sintáctico

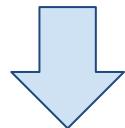
Estudia las reglas para combinar palabras en sintagmas y oraciones.

- Tokenización
- **División en oraciones**
- Análisis sintáctico (superficial, profundo, de dependencias)

Nivel Sintáctico - División en oraciones

- Sentence splitting: dividir el texto en oraciones.

Carolina es Ingeniera Informática. Ella estudió en la UC3M.



Carolina es Ingeniera Informática.

Ella estudió en la UC3M.

Nivel Sintáctico - División en oraciones

A veces es dependiente de la tokenización

*La página personal de la Dra. Segura es
hulat.inf.uc3m.es/nosotros/miembros/isegura*

Las acciones de Repsol S.A se han incrementado un 3.5%.

Nivel Sintáctico - División en oraciones

Según el tipo de texto (redes sociales, notas clínicas) , la tarea extremadamente difícil!!!



Nivel sintáctico - división oraciones

- Enfoques:
 - expresiones regulares, listas de acrónimos, lista de excepciones (por ejemplo, nombres propios)
 - enfoques basados en algoritmos (estadísticos y aprendizaje automático) para los casos más complejos.

Nivel Sintáctico

Estudia las reglas para combinar palabras en sintagmas y oraciones.

- Tokenización
- División en oraciones
- **Análisis sintáctico (superficial, profundo, de dependencias)**

Nivel Sintáctico - Análisis Sintáctico

- Análisis sintáctico superficial, shallow parsing, chunking

He reckons the current account deficit will narrow to only # 1.8 billion in September .



[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Nivel Sintáctico - Análisis Sintáctico

- Análisis sintáctico (superficial, profundo, de dependencias)
- Basado en gramáticas.

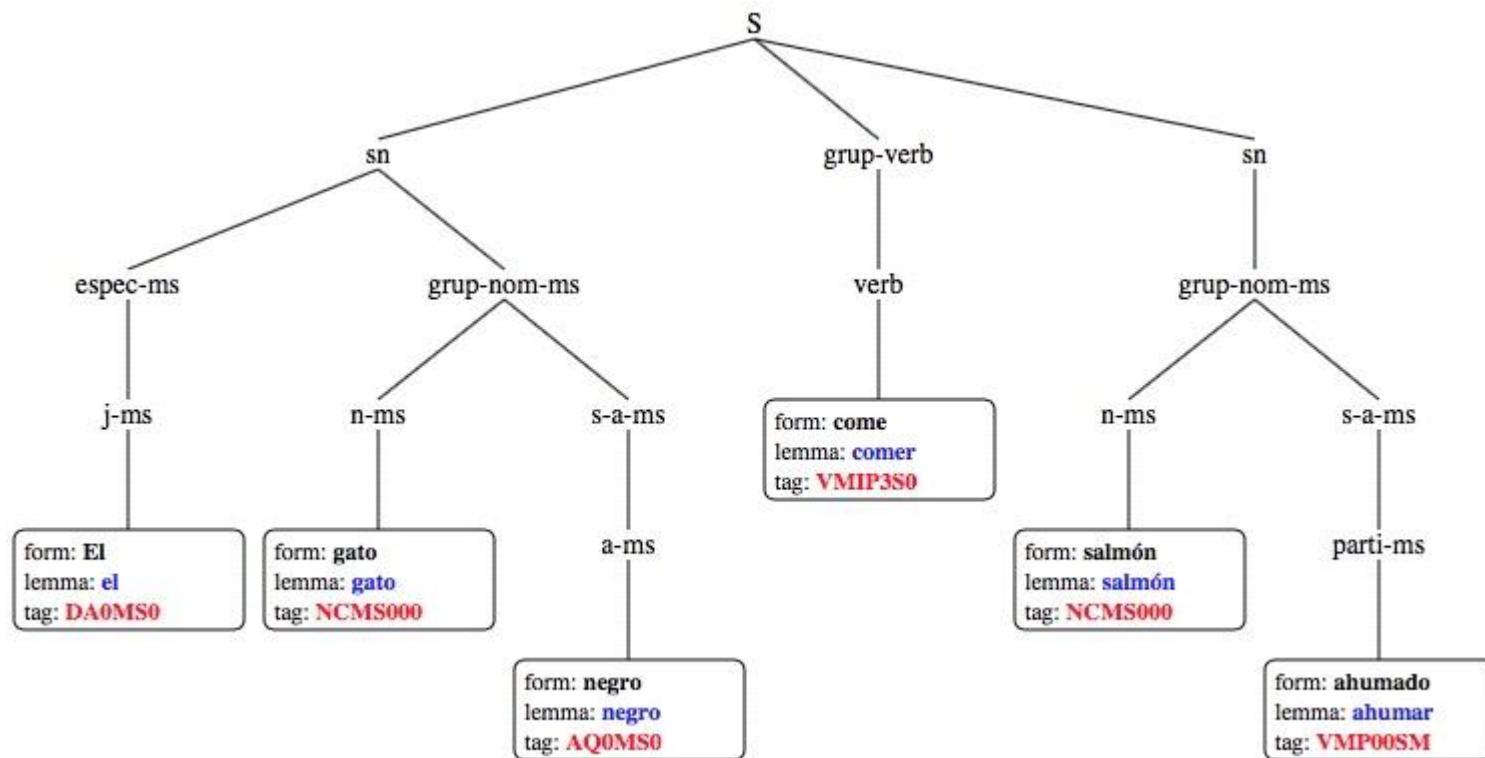
Oración -> SN SV
SN -> (Det) Nombre Adj.
SV -> Verbo SN

SN= Sintagma nominal,
SV = Sintagma verbal

S=El gato negro come salmón ahumado

Nivel Sintáctico - Análisis Sintáctico

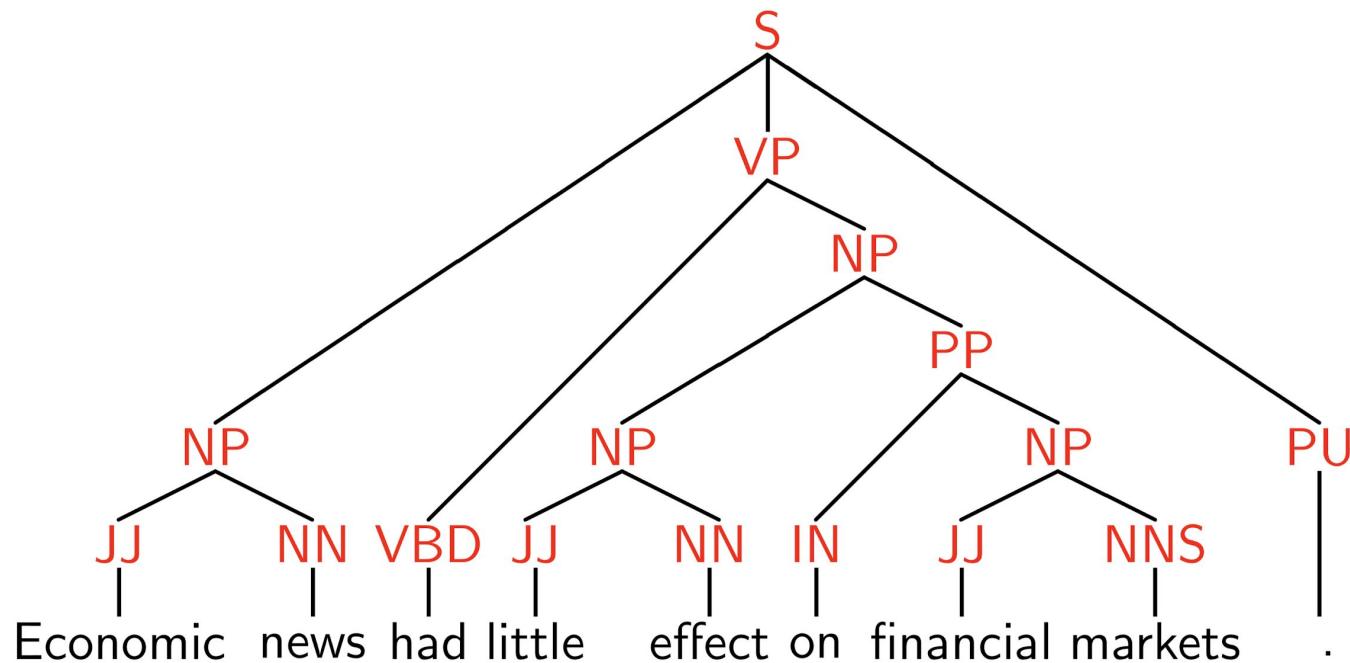
El gato negro come salmón ahumado



Padró, L. (2016). Demonstration. *FreeLing 4.0. An open-source suite of language analyzers*. Barcelona: TALP - Tecnologies i Aplicacions del Llenguatge i de la Parla, Universitat Politècnica de Catalunya. Consultado en <http://nlp.lsi.upc.edu/freeling/demo/demo.php>

Nivel Sintáctico - Análisis Sintáctico

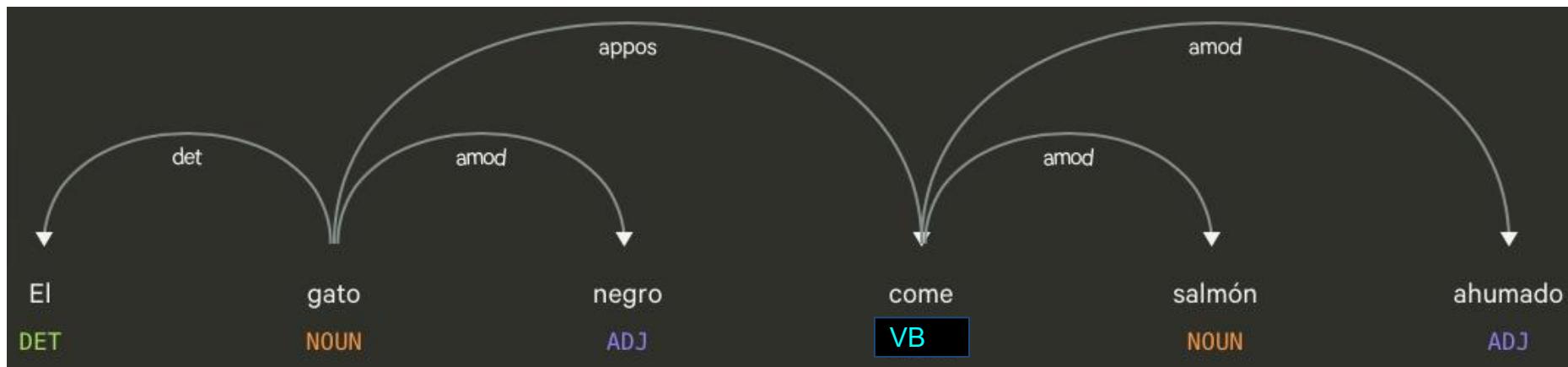
Economic news had little effect on financial markets.



Nivel Sintáctico - Análisis Sintáctico

Análisis de dependencias

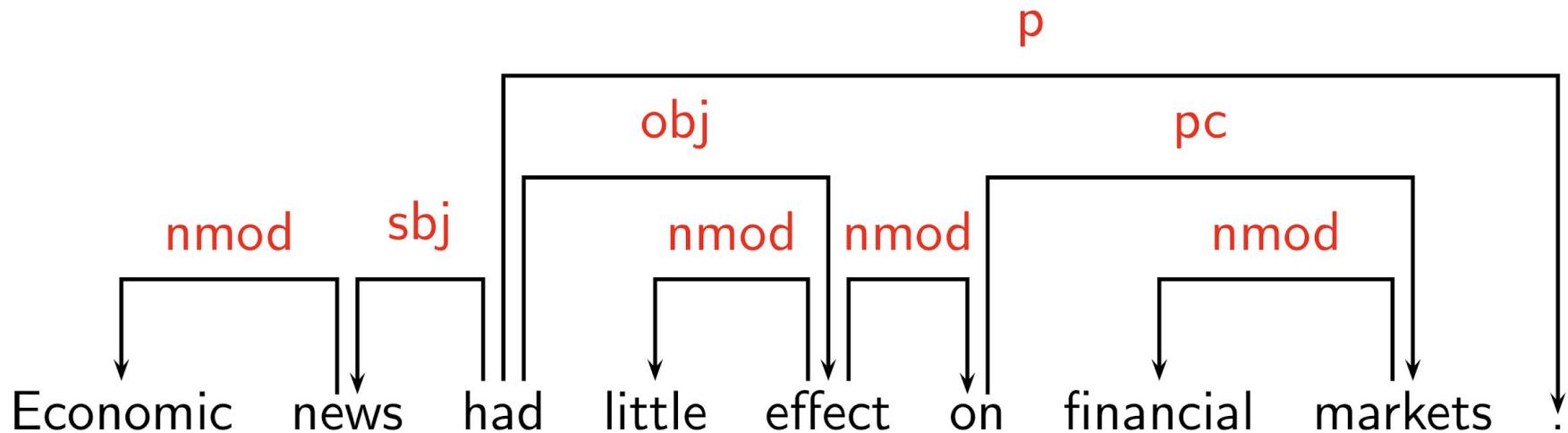
El gato negro come salmón ahumado



Análisis de dependencias. det: *determinant*; amod: *adjectival modifier*; appos: *appositional modifier*; VB *verb*.

Nivel Sintáctico - Análisis Sintáctico

Análisis de dependencias



Ambigüedad en el nivel sintáctico

- *María guardo las revistas que Paco dejó bajo la mesa* - ¿Paco había dejado las revistas bajo la mesa y María las guardo, o María guardo las revistas bajo la mesa?
- *La matanza de cazadores fue espantosa* - ¿los cazadores fueron los verdugos o las víctimas?
- *La hermana de Lourdes, a quien conocí ayer, vino a verme.* - ¿A quién conocí ayer: a Lourdes o a su hermana?
- *María encontró desquiciada a Clara.* ¿Quién estaba desquiciada María o Clara?
- *Hablé con el dueño del gato que estaba jugando en la cancha* - ¿Quién estaba jugando el gato o el dueño?

Nivel Semántico y Pragmático

Nivel léxico-semántico: estudio del significado de las expresiones del lenguaje.

Nivel Pragmático: estudio del lenguaje en el contexto (intención real del mensaje teniendo en cuenta el contexto y el mundo).

- *Te espero mañana donde siempre.*
- *Por qué no te callas???* (*Rey Juan Carlos I / Chavez*)
- *No estoy, me compro un desierto y me fui a barrerlo!.*

Ambigüedad en el nivel léxico - semántico

- *Juan aquí no pinta nada más* - se puede referir Juan sobra o que Juan ya no tiene que pintar paredes.
- *Han puesto un banco nuevo en la plaza* - banco de sentarse o banco financiero.
- *Pedro quiere pelearse con un francés* - ¿un francés concreto o le sirve cualquier francés para pelearse?
- *Juan come mucho* - ¿Juan come mucha cantidad o de forma frecuente?
- *Las niñas comieron galletas. Son buenas* - ¿las niñas o las galletas?
-

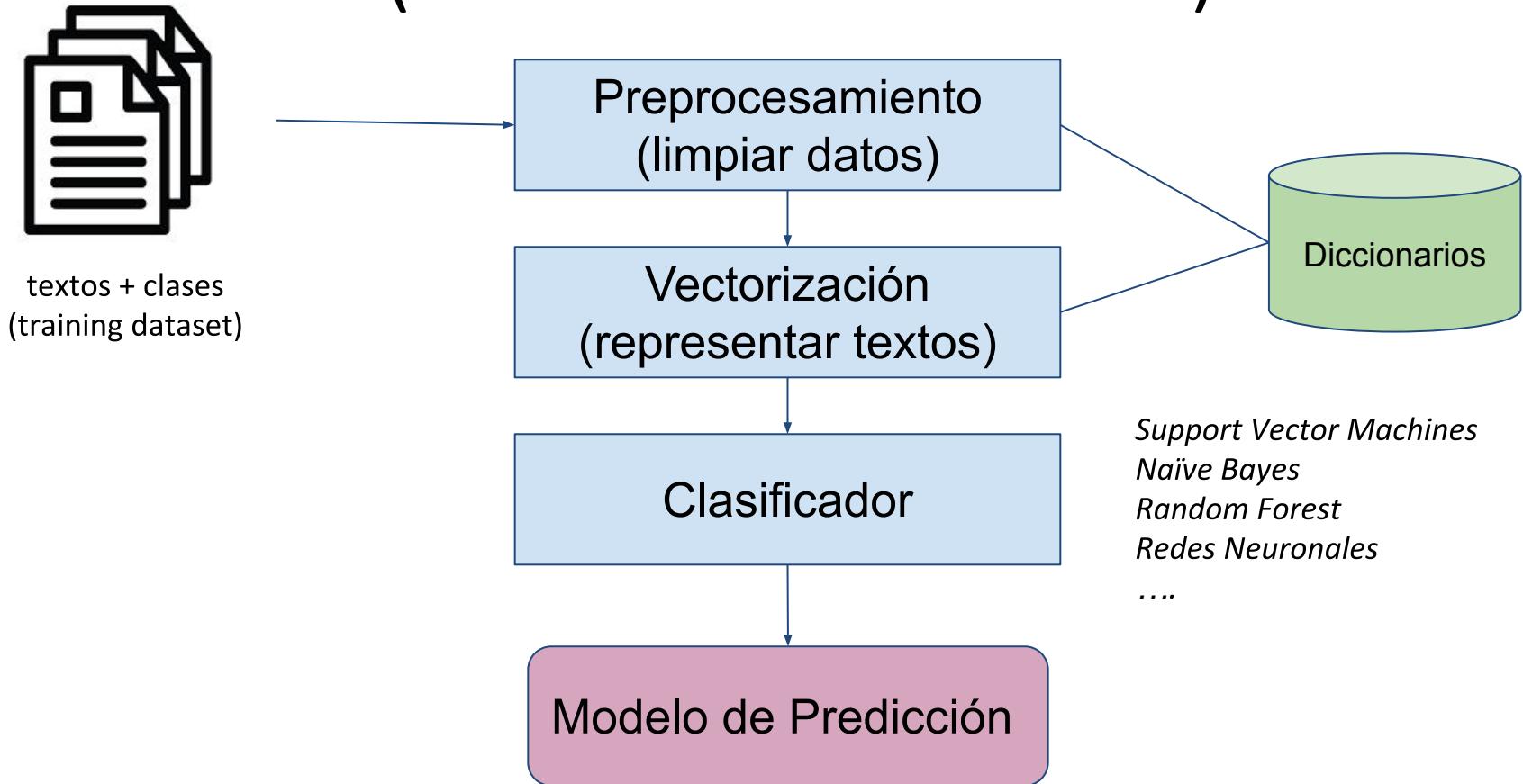
Ambigüedad en el nivel pragmático

- *Juan golpeó el armario con el bastón y lo rompió* - ¿el bastón o el armario?. Nos falta más conocimiento para saber qué ocurrió.
- *El abogado salió a comer con la esposa* - ¿Con qué esposa salió a comer, con su esposa o con la esposa de algún cliente?.
- *Ella ya ha tomado suficiente.* ¿Comida, bebida o ejercicio?

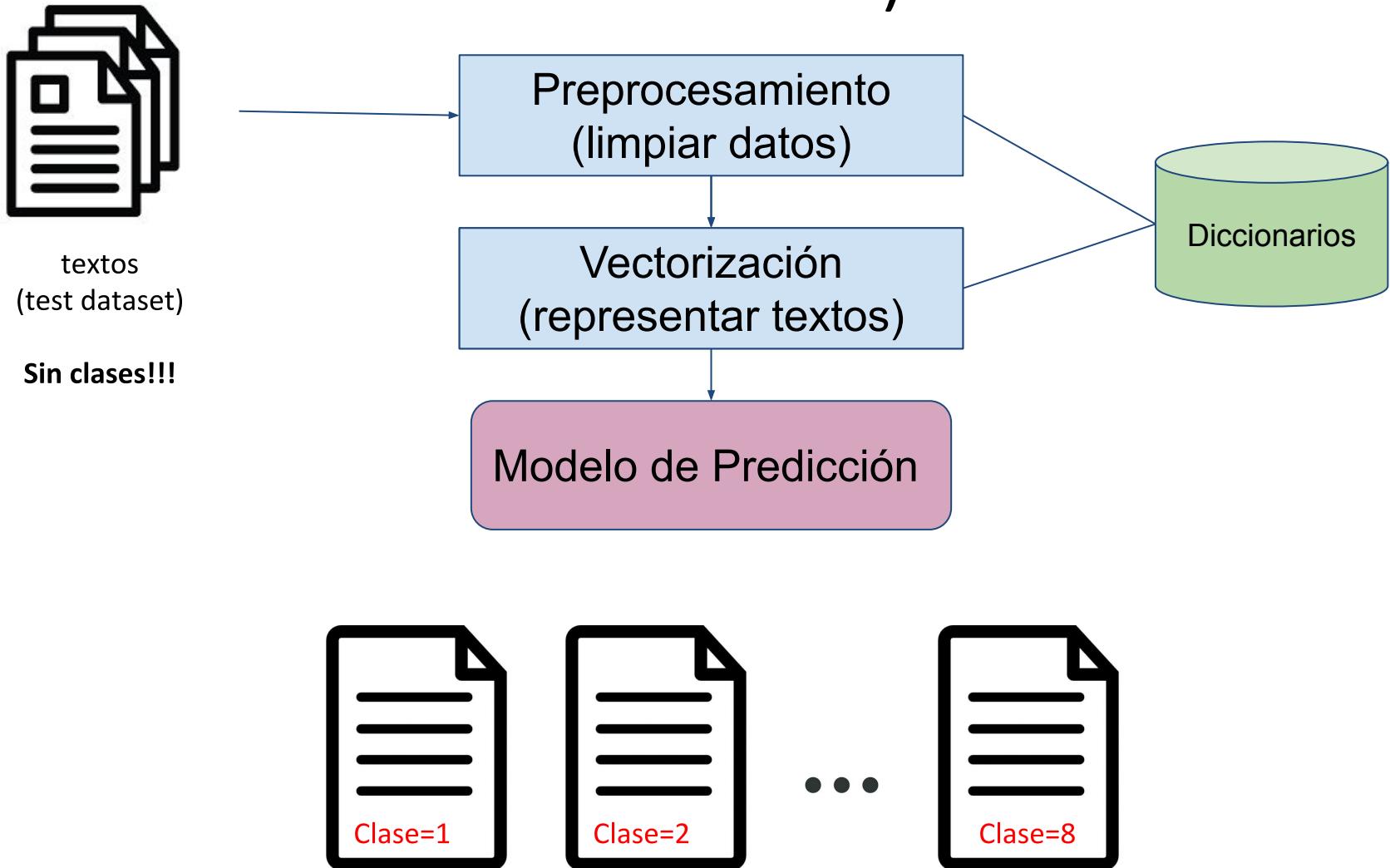
¿Cómo es un sistema de PLN?

- Pipeline (secuencia) de componentes.
- Cada componente aborda algún nivel del lenguaje o parte de él.
- Dependiendo de la aplicación final, el sistema estará compuesto por distintos componentes y abordará distintos niveles del lenguaje.

Arquitectura Clasificación de Textos (fase de entrenamiento)

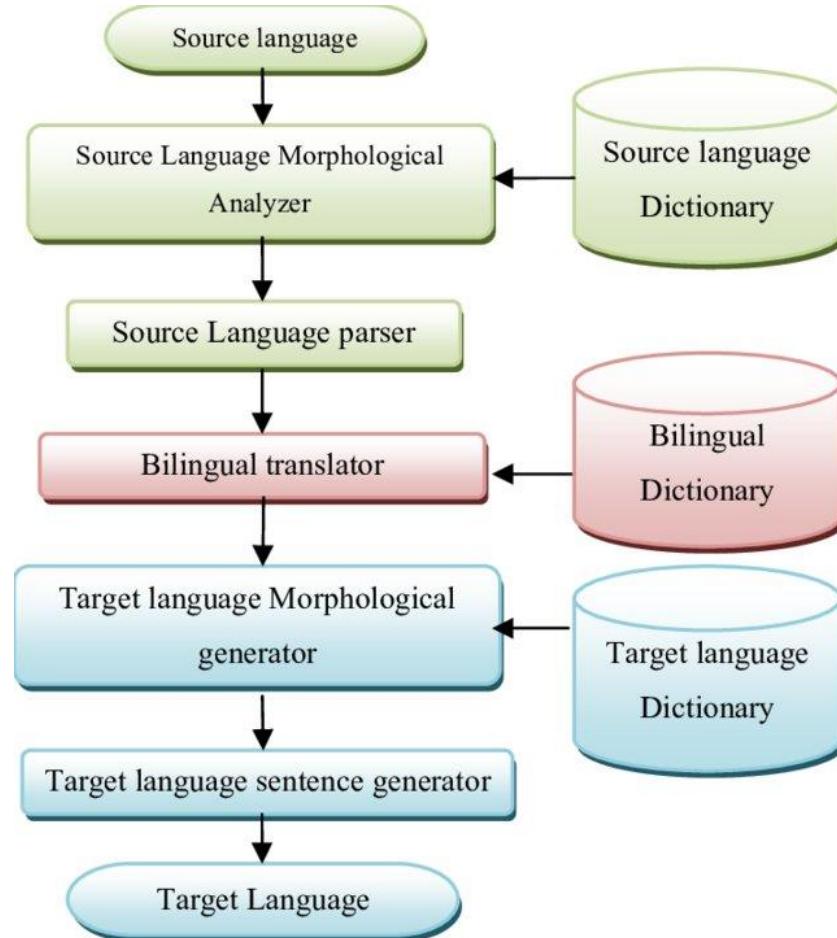


Arquitectura Clasificación de textos (fase de evaluación)



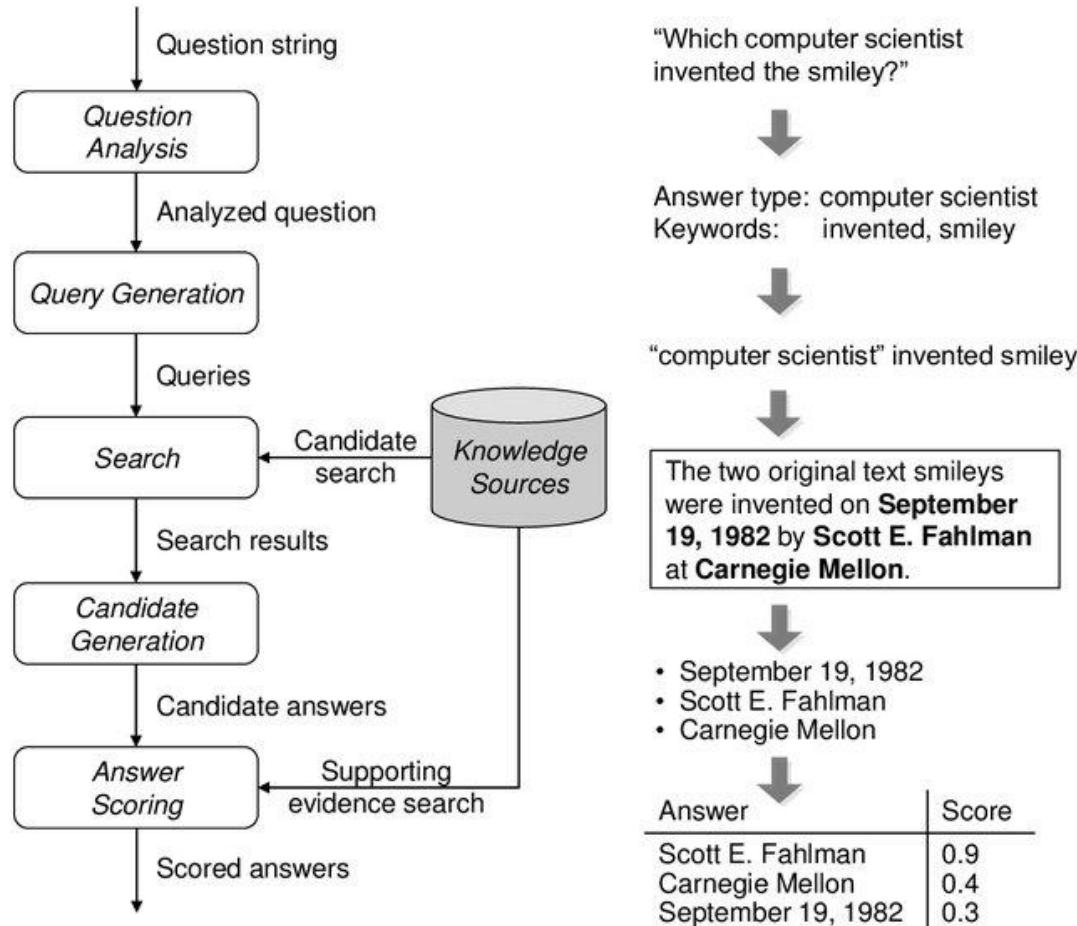
Predicción de las clases para cada texto del test dataset

Arquitectura de un sistema Traducción automática



Hettige, B., & Karunanananda, A. S. (2011, September). Computational model of grammar for English to Sinhala Machine Translation. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 26-31). IEEE.

Sistema Búsqueda de Respuestas (Question Answering)



Schlaefer, N., Chu-Carroll, J., Nyberg, E., Fan, J., Zadrozny, W., & Ferrucci, D. (2011, October). Statistical source expansion for question answering. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 345-354). ACM.

Conclusión

- Ingente cantidad de datos, gran parte en formato no estructurado => PLN como solución para el análisis de datos no estructurados.
- Aplicación en diferentes dominios (financiero, turismo, biomedicina, clínico).
- Conocimiento estructurado en niveles.
- Ambigüedad principal reto a resolver. Está presente en todos los niveles.